PERICLES - Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics

[Digital Preservation]

DELIVERABLE 4.3

CONTENT SEMANTICS AND USE CONTEXT ANALYSIS TECHNIQUES

GRANT AGREEMENT: 601138

SCHEME FP7 ICT 2011.4.3

Start date of project: 1 February 2013

Duration: 48 months

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	PUBLIC	X
PP	Restricted to other PROGRAMME PARTICIPANTS (including the Commission Services)	
RE	RESTRICTED to a group specified by the consortium (including the Commission Services)	
CO	CONFIDENTIAL only for members of the consortium (including the Commission Services)	

Revision History

V #	Date	Description / Reason of change	Author
V0.1	23.10.15	Outline and first draft	CERTH
V0.2	23.11.15	Additions from all partners	CERTH, KCL, SpaceApps, HB
V0.3	04.12.15	Refinements to chapters 3, 4	CERTH, KCL, SpaceApps, HB
V0.4	09.12.15	Improvements to chapters 2, 3, 4	CERTH, KCL, SpaceApps, HB
V0.5	21.12.15	First round of internal reviewing	KCL, CERTH
V0.6	04.01.16	Restructuring of Chapter 2, refinements to the rest of the document	CERTH, HB
V0.7	25.01.16	Second round of internal reviewing	KCL, DOT, CERTH
V1.0	28.01.16	Final version of the deliverable	CERTH, KCL

Authors and Contributors

Authors

Partner	Name
CERTH	A. Maronidis, E. Chatzilari, E. Kontopoulos, S. Nikolopoulos, M. Riga, P. Mitzias
HB	S. DaránYi, P. Wittek
KCL	A. Gill, E. Tonkin
SpaceApps	D. De Weerdt
ULIV	F. Corubolo

Contributors

Partner	Name
KCL	S. Waddington, C. Sauter

Reviewers

Partner	Name
KCL	S. Waddington
DOT	S. Tekes

Table of Contents

Glossary

Abbreviation / Acronym	Meaning
AL	Active Learning
CM	Classical (Newtonian) Mechanics
CNN	Convolutional Neural Networks
CS	Compressed Sensing
CSI	Content Semantic Information
DO	Digital Object
DP	Digital Preservation
GDS	Generalized Differential Sparsity
GI	Gini Index
LRM	Linked Resource Model
LTDP	Long Term Digital Preservation
mAP	mean Average Precision
PCA	Principal Component Analysis
PCS	Product Compressed Sensing
PET2LRM	PET (PERICLES Extraction Tool)-to-LRM (Linked Resource Model) conversion scheme
PQ	Product Quantization
SALIC	Social Active Learning for Image Classification
QM	Quantum Mechanics
SIFT	Scale-Invariant Feature Transform
SCP	Significant Content Properties
SP	Significant Properties

1. Executive Summary

The current deliverable summarises the work conducted within task T4.3 of WP4, focusing on the extraction and the subsequent analysis of semantic information from digital content, which is imperative for its preservability. More specifically, the deliverable defines content semantic information from a visual and textual perspective, explains how this information can be exploited in long-term digital preservation and proposes novel approaches for extracting this information in a scalable manner.

Additionally, the deliverable discusses novel techniques for retrieving and analysing the context of use of digital objects. Although this topic has not been extensively studied by existing literature, we believe use context is vital in augmenting the semantic information and maintaining the usability and preservability of the digital objects, as well as their ability to be accurately interpreted as initially intended.

The main contributions of the deliverable are the following:

- An original method for measuring data sparsity that extends the limits of existing methodologies for compression in the context of content decomposition and feature extraction.
- A novel field theory of semantic content for modelling the evolution of document content over time.
- Two newly introduced approaches for semantic concept detection and content classification that efficiently tackle scalability issues.
- A prototype tool that scalably processes semantic media content based on vector fields, which paves the way for the exploration of quantum-like content behaviour in digital collections.
- Two novel approaches for analysing source text documents relevant to the two case studies, along with respective tools for populating the domain ontologies with instances based on the extracted information.
- An ontology-based approach for representing the context of use of digital objects based on LRM constructs.
- A prototype method for extracting and analysing use context information, which is based on two existing PERICLES outputs, PET and its PET2LRM plugin.
- Respective implementations and experimental results that validate all the above.

All these contributions are tightly interlinked with the endeavours in the other PERICLES work packages: WP2 supplies the use cases and sample datasets for validating our proposed approaches, WP3 provides the models (LRM and Digital Ecosystem models) that form the basis for our semantic and use context representations, WP5 provides the practical application of the technologies developed to preservation processes, while the tools and algorithms presented in this deliverable can be used in combination with test scenarios, which will be part of the WP6 test beds.

2. Introduction & Rationale

Digital Objects (DOs) are typically accompanied by a set of metadata describing their technical characteristics (e.g. file format, aspect ratio, frame rate, etc.) and provenance (e.g. creator, date and place of creation, etc.). An additional type of useful metadata is context information, i.e. any information coming from the environment that can be used to characterise the situation the DO is in. Such types of metadata, i.e. technical characteristics, provenance and context, play a key role in ensuring access to the DOs and have already been at the centre of attention in PERICLES, mostly in WPs 2, 3 and 4. However, although they provide significant information about a DO and its history, they do not reveal much about its **semantic content** (i.e. the notions, meanings, topics and themes encompassed by the object) and **context of use** (i.e. information relevant to the context of use of the DO). This is exactly the focus of the current deliverable, i.e. **to propose novel methodologies for extracting and analysing the semantic content** of DOs and to also move one step further into **proposing a framework for gathering and analysing the context of use** of the digital content.

We note that contextual data may be usefully collected at the individual object level, as well as at collection level, or as data relating to the host institution, since these levels may also play a role in understanding and interpreting such DOs. Policies and strategy documents are examples of data relating to the host institution, as are materials authored by other groups (government, media, focus groups, public documents and so forth), which shed light on various aspects of the cultural, social and financial environment of the organisation and its activities. We believe that the contextual information relating to digital content plays a significant role in maintaining the usability and preservability of the DOs, as well as in enhancing their ability to be interpreted as it was initially intended when they were first created.

2.1. Context of this Deliverable

This deliverable summarises the work of task T4.3 and documents our prototype frameworks for the retrieval and semantic analysis of digital content and its use context. It discusses what content semantics from a visual and textual perspective is, how this semantics could be extracted, modelled and analysed. It explores how the representation of use context can be deployed in augmenting the semantic information and how all of these investigations relate to the other relevant research activities within PERICLES.

2.1.1. What to expect from this Document

As already mentioned, this deliverable looks at two kinds of context-dependent content, **image** and **text semantics**, although our findings are generalizable to other document modalities which are represented by their features (e.g. video, audio, etc.), too. Basically, the following three key topics are discussed in the document:

- **Content decomposition and feature extraction**: The deliverable discusses the role of content decomposition and feature extraction methods in representing a DO in a form that allows for the automatic extraction of semantic information. As the above methods produce massive data volumes that are impossible to handle, a compression step that allows for a subsequent reconstruction of the original content is proposed. In order to be able to recover as much semantic information as possible during reconstruction, a novel reconstruction method is proposed.
- **Semantic concept detection and content classification**: In this direction, the deliverable discusses existing methodologies for semantic information extraction, observing the activity and its rapidly increasing relevance that underpins the importance attached to this field of research.

Going one step beyond the state-of-the-art, the deliverable proposes novel methods for semantic information extraction based on machine learning techniques. As scalability proves to be a key requirement in the whole process, we propose techniques that are able to cope with the large-scale aspect of the data.

- **Extraction and analysis of use context information**: Finally, the deliverable discusses the proposed approaches for representing, extracting and analysing use context information, i.e. any information that is relevant to the contexts of use of a DO. For the representation of use context, we discuss the relevant structures adopted by the domain ontologies, while for the extraction and analysis of this information, we are deploying the Pericles Extraction Tool (PET) and the PET2LRM conversion scheme, which were presented in deliverables D4.1 and D4.2, respectively.

2.1.2. Relation to other Work Packages

The work presented in this deliverable is strongly linked to other PERICLES WPs as follows:

- Ontology population activities (see Section 4.2) operate on the domain ontologies (WP2), which in turn are based on LRM representations (WP3). In essence, the LRM is a change management formalism (represented as an ontology), used by the domain ontologies to model the Digital Preservation (DP) risks for the two domains considered by the project: Space Science and Art & Media. Consequently, delivering mechanisms for populating the domain ontologies (in a user-driven, semi-automatic fashion) and for enriching their set of DO instances offers non-trivial added value to the project outcomes.
- The same holds for the use context extraction and analysis activities (see Chapter 5), which again operate on the domain ontologies, as specified in the DoW. The link to WP3 and the LRM here is stronger, since the “Dependency” construct defined in the LRM, and, more specifically, the enclosed notions of “Intention” and “Specification”, as well as the extensions/ specialisations of these concepts, form the basis for the adopted representation of use context.
- The work on semantic and use context analysis provides a set of methods and components, which contribute to the work on management of change in WP5. Specifically, subtask T5.3.2 is analysing how practical tools for curators can be provided to manage risks of change in semantics and user communities in cultural heritage collections. T5.4 is investigating ongoing appraisal of digital content in the face of evolving user communities, which can also result in changing requirements for metadata. Our approach to these two tasks is described in D5.2 [PERICLES D5.2, 2015].
- The feature extraction and classification algorithms developed and presented in this deliverable (see Chapters 3 and 4) have been applied to one of Tate’s visual repositories (WP2) in retrieving content based on high-level concept queries. In a similar manner, the above methods can be adopted by any memory organisations (interested in or already deploying DP methodologies) that have large-scale visual repositories; towards this end we have developed an open source library performing concept detection from visual content.
- All the tools and algorithms presented in this deliverable can be used standalone, but also in combination with test scenarios, which will be part of the WP6 test beds running within the integration framework.

2.1.3. Relation to other WP4 Tasks

Besides the interconnections with other WPs, the work presented here is tightly linked to the other WP4 tasks as well. In essence, the analysis activities in T4.3 are following the extraction and encapsulation activities from T4.1 and T4.2, and will, in turn, be followed by the modelling activities in T4.4 and the interpretation and reasoning activities in T4.5. Consequently, the modelling activities are connecting analysis with interpretation. More specifically:

- The extraction and analysis of use context information (see Chapter 5) currently relies on methodologies and tools developed within T4.1 (PET) and T4.2 (PET2LRM) - the latter tool is still under development (release is due M44) and will contribute in retrieving use context information for analysis within the tasks relevant to contextualised content semantics (T4.4) and contextualised content interpretation (T4.5).
- Our proposed methodologies for semantic extraction can provide a set of use context metadata, which, aggregated with the other types of metadata, can serve as input to PeriCAT (see D4.2).
- The definitions and state-of-the-art surveys for content semantics and (use) context as well as the respective formalisms and representations are based on work being conducted within T4.4 "*Modelling content semantics, interpretations and context*", which will be presented in full in the upcoming deliverable D4.4 due M40.
- Finally, this deliverable includes some key considerations about a new conceptual framework for DP that asks about the possibly quantum-like[1] behaviour of objects and features in evolving digital collections. Because our respective findings will be reported within the context of T4.5 in D4.5 (due M44), it is sufficient here to refer to two immediate connections: (a) the active learning method for image content analysis that holds the promise of further acceleration and scalability in future computations; (b) another method inspired by mathematics and quantum mechanics, and one for which we developed a tool, which investigates the very foundations of machine learning, namely those correlation types underlying statistical groupings of DOs and features.

2.2. Document Structure

The structure of this document follows the overview outlined in Section 2.1.1 and is as follows:

- **Chapter 3 - Content Decomposition and Feature Extraction**: In this chapter, the state of the art approaches for image representation are described. More specifically, two basic image representation schemes are analysed; a) a pipeline approach consisting of a key-point detection, a feature extraction and a feature encoding part, while indicative state-of-the-art methods pertaining to each of the three above parts are presented and b) the popular and more recent features based on Convolutional Neural Networks (CNNs). Moreover, as the above pipeline approach leads to massive data, the role of data compression and reconstruction in the effective extraction of semantic information (see Chapter 4) is also discussed. As the correct functionality of compression is strongly dependent on the sparsity of the data, we propose a novel approach for measuring sparsity that extends the limits of the adopted methodology for compression.
- **Chapter 4 - Semantic Concept Detection and Content Classification**: This chapter is focused on methods that opt to render the approaches presented in the previous chapter scalable to massive datasets. Towards this direction, we present two methods, called SALIC and PCS, that attempt to tackle the scalability issue in two ways respectively; 1) by reducing the number of required training instances (SALIC) and 2) by reducing the dimensionality of the features (PCS). More specifically, SALIC gathers training data without requiring significant annotation efforts, while at the same time minimises the number of the required training instances and increases the performance of the classification models by utilising a smart sampling approach. PCS, on the other hand, is a very fast method for dimensionality reduction. With regards to semantic information extraction from text-based content, the chapter also presents our approaches for analysing source text documents relevant to the two case studies and using the extracted information for populating the developed domain ontologies with instances.

[1] Recent research on cognitive systems indicates non-trivial similarities between quantum and classical physics, hence we believe that quantum-like modelling in linguistic, information retrieval, and digital library research is relevant and important in spite of being still exploratory.

- **Chapter 5 - Extraction and Analysis of Use Context Information**: This chapter focuses on a specific type of context, “use context”, which refers to information related to contexts of use of the DO, and discusses the adopted approaches for extracting and analysing use context information, in order to address issues like e.g. variations of DO interpretation, and to derive meaningful correlation links among content objects and use contexts. For representing use context, we deploy appropriate structures adopted by the domain ontologies. For extracting and analysing this information, on the other hand, we are using a core PERICLES tool, PET, along with its PET2LRM plugin that allows converting PET JSON output into LRM snippets.
- **Chapter 6 - Conclusions & Future Work**: Finally, the deliverable concludes with some final remarks and an account of potentially interesting directions for future work, with regards to each of the key topics discussed in the previous chapters.

3. Content Decomposition and Feature Extraction

LTDP is facing the complex challenge of three kinds of dynamics affecting our cultural heritage: unavoidable changes in technology, language and cultural values [Schlieder, 2010]. With new, bold methodologies heralding new ways of thinking [Goldman et al., 2013; Grass et al., 2015], we are addressing the problem of modelling content dynamics as of core significance to PERICLES. In this chapter, we describe the role of feature extraction methods in representing a DO based on the decomposition of its content. We also discuss the role of data compression in the effective extraction of semantic information from the original content of a DO. More specifically, we analyse the assumptions that guarantee the correct operation of the compression process and towards this end we propose a novel methodology that extends the functionality limits of the adopted methods in PERICLES for compression. The methods deployed here are used for both image and text analysis.

3.1. Visual Content Decomposition and Feature Extraction

Mapping visual content to semantics has been an important research direction for many decades. A typical pipeline includes the representation of the visual content's appearance in a form amenable to machine processing and then applying machine learning techniques to the representations. Initially, researchers tried to capture the color, texture and shape of an object in order to describe it [Bober, 2001; Manjunath et al., 2001]. More recently, the SIFT-alike (scale-invariant feature transform) features became the state-of-the-art approach to represent images [Chatfield et al., 2011]. The typical pipeline for image representation, which has been adopted by PERICLES, consists mainly of three steps: (a) Key-point detection, (b) Feature descriptor extraction, and, (c) Encoding of the features into a single vector representation of the image. In the beginning, key-points are detected from an image, at which local feature descriptors are extracted. A feature encoding algorithm is then used to represent an image with a single vector.

Key-point detection refers to the detection of salient points in an image that can robustly characterize the latter. They are usually points of high interest representing important aspects of the image. Visual recognition requires repeatable key-points under several transformations such as rotation, scale changes, translation and lighting variations, whilst keeping the detection time to the minimum. Key-point detectors differentiate on invariance in image transformations, computational time and repeatability. Scale-invariant feature transform (SIFT) [Lowe, 2004], Harris Corners [Harris & Stephens, 1998], maximally stable extremal regions (MSER) [Matas et al., 2004] are some examples. Mikolajczyk and Schmid (2005) review several key-point detectors on a benchmark dataset.

Feature extraction attempts to describe the surrounding environment of each key-point, so that it captures characteristic and indicative information of its visual content. Each key-point that was previously detected is represented as an n-dimensional feature vector (descriptor). The SIFT descriptor [Lowe, 2004], the very first algorithm that actually brought this image representation pipeline in light, is still the most popular algorithm in this area. However, SIFT is a very computationally expensive algorithm and thus cannot be easily used in real time applications. Many approaches have attempted to alleviate this problem by proposing alterations, which as expected led to performance loss. The most popular of these variations is Speeded Up Robust Features (SURF) [Bay et al., 2008]. Several colour variations of SIFT are evaluated in [Van De Sande et al., 2010].

Feature encoding transforms the set of n-dimensional vectors obtained via key-point detection and feature extraction into a single vector. The most popular method, borrowed from text retrieval, is the "bag of visual words" [Lazebnik et al., 2006]. A vocabulary of visual words is constructed from a large independent training set of images by clustering the descriptors that are extracted from them using an algorithm such as k-means to n clusters/visual words. Ideally, these visual words should carry enough information to distinguish images into categories. Afterwards, the feature encoding algorithm is applied. The baseline of the bag of visual words approach employs the vector quantization idea and is mostly known as the "hard assignment" method. Key-points are assigned to the nearest visual words in the vocabulary and the histogram is computed by adding 1s to the corresponding words. Later, the soft or Kernel codebook assignment method was proposed and exhibited better performance [van Gemert et al., 2008; van Gemert et al., 2010]. The authors introduced the visual word ambiguity term and for every key-point, instead of adding 1s to only the nearest visual word they added its kernel distance to each visual word. In the last couple of years many approaches were proposed in order to increase the performance of the soft assignment encoding method. The most popular of them include the Vector of Locally Aggregated Descriptors (VLAD) [Jégou et al., 2010], the Fisher vector encoding [Perronnin et al., 2010], the super vector encoding [Zhou et al., 2010] and the Locality-constrained linear (LLC) encoding [Wang et al., 2010]. An extensive empirical study on encoding methods can be found in [Chatfield et al., 2011].

Lately, features based on **deep learning techniques** have shown remarkable performance[2] and have been established as the main feature extraction technique. Although the basic theory for deep learning has been around for some time and has been also investigated in the past, this rather unexpected breakthrough was mainly attributed to: (a) the large volumes of training data that became available through ImageNet[3], and b) the high-processing capabilities of modern GPUs that allowed for learning very deep networks in an acceptable amount of time. With the impressive results of Convolutional Neural Networks (CNNs) in both image annotation and object detection [Krizhevsky et al., 2012], many researchers have investigated their potential to facilitate various computer vision tasks. In [Chatfield et al. (2014)], the authors present an extensive study and comparison between features originating from CNNs and SIFT-alike features followed by encoding algorithms (e.g. Bag-of-Words, Fisher encoding, etc.). Similarly, in [Hariharan et al. (2014)], the authors show that the parameters of CNNs can be learnt on independent large-scale annotated datasets (such as ImageNet) and can be efficiently transferred to other visual recognition tasks with limited amount of training data (i.e. object and action detection). Towards the same objective of object detection, Oquab et al. (2014a) present a method that is also based on CNNs, which simultaneously segments and detects objects in images. Finally, based on weak but noise-free annotations, Oquab et al. (2014b) present a weakly supervised CNN for object recognition that does not rely on detailed object annotations and shows that it can perform equally well when strong annotations are present.

Finally, after representing each image with a single feature vector, a model that learns the correspondence between image labels and features needs to be trained. One way to accomplish this is by using probabilistic methods which try to find the joint distribution between labels and features (e.g. Bayesian Networks [Domingos & Pazzani, 1997]) or the conditional distribution of the labels given the features (e.g. conditional random fields [He et al., 2004]). There are also the tree decision algorithms [Breiman et al., 1984], which attempt to map observations about an example to conclusions about the example's true class. Random forests [Breiman, 2001] is an example of such an algorithm, which constructs a number of random decision trees in a controlled way in order to obtain better predictive performance and generalization. Neural networks [Egmont-Petersen et al., 2002]

[2] http://image-net.org/challenges/LSVRC/2012/results.html
[3] http://www.image-net.org/ (see also [Deng et al., 2009a])

are inspired by the structure and functionalities of human neural networks and attempt to capture the structure in the unknown joint probability distribution between observed variables. Finally, there exist algorithms that attempt to split the feature space so that the different classes are separated. Logistic regression [Hosmer & Lemeshow, 2004] and Support Vector Machines (SVMs) [Cortes & Vapnik, 1995] are the most popular in this category.

The process of learning any classifier typically consists of feeding the machine learning algorithm with positive and negative instances of the targeted concept in order to train a classification model. Although this process usually requires manual annotation on unlabeled data, which is a time consuming task, recently there have been approaches for automatically gathering training content, requiring only minimal amounts of data to be manually annotated [Chatzilari et al., 2014; Li et al., 2013; Vijayanarasimhan & Grauman, 2014].

3.2. Compression and Reconstruction of Big Data

The feature extraction and encoding techniques presented in the previous subsection, often result to very big data streams representing the original DOs. Thus, a compression/dimensionality-reduction step prior to the application of machine learning techniques is imperative, in order to be able to treat the DO in a computationally efficient manner. However, the compression of the original DO must allow for the efficient extraction of semantic information. For this purpose, it is equally important to be able at any time to reconstruct the original DO from its compressed version. Therefore, in this deliverable we are considering both the compression and reconstruction processes.

In this subsection, we first review the state-of-the-art in dimensionality reduction and data compression putting specific emphasis on **Compressed Sensing (CS)**, which has recently emerged as a powerful method in signal compression-reconstruction. In addition, we illustrate the significant role of data sparsity and/or compressibility in the correct operation of CS and the optimal reconstruction of the original data. Based on this, we review a number of state-of-the-art techniques that have been proposed for estimating/calculating data sparsity and we propose our novel **Generalized Differential Sparsity (GDS)** framework. The effectiveness of our approach in reconstructing the original DO from its CS-compressed version is proven through an experimental study.

3.2.1. Dimensionality Reduction

As enormous volumes of data are being created every day, great interest has been placed on the theoretical and practical aspects of extracting knowledge from massive data sets [Bacardit & Llorà, 2013; Dean & Ghemawat, 2008]. For dealing with the new challenges, a number of “big-data aware” techniques have recently gained attention, including incremental updating, smart sampling, and parallel and distributed architectures [Sánchez & Perronnin, 2011]. Although these techniques have a strong potential in big data processing, most of them benefit from a dimensionality reduction step of the data. A plethora of methodologies have recently been proposed for reducing the dimensionality of large-scale data. These methodologies can be classified into three main categories.

1. **Methods based on statistics and information theory**: This category, among others, includes Vector Quantization (VQ) [Gray & Neuhoff, 1998] and Principal Component Analysis (PCA) [Jolliffe, 2002]. In the same direction, new scalable techniques, such as Product Quantisation (PQ) [Jégou et al., 2011], have also been developed. PQ decomposes a vector space into a Cartesian product of quantised subspaces for constructing short codes representing high-dimensional vectors. In this vein and using the Vector of Locally Aggregated Descriptors (VLAD), Jégou et al. (2010) employ PQ for performing nearest neighbor search during the indexing process of very large databases for retrieving the most similar images to a query image. Moreover, within the vector quantization context, effort has recently been invested on the

optimisation of the kernel K-Means. For instance, Yu et al. (2012) propose a clustering algorithm that models multiple information sources as kernel matrices.

2. **Methods based on dictionaries:** Here, a vector is represented as a linear combination of the atoms of a dictionary. For instance, sparse representations with over-complete dictionaries have been applied to image denoising [Elad & Aharon, 2006]. Utilizing the K-SVD technique, the authors train dictionaries based on a set of high-quality image patches or based on the patches of the noisy image itself. The image is iteratively de-noised through a sparse coding and a dictionary update stage. Based on a similar approach, K-SVD has been utilized for the restoration of images and videos in [Mairal et al., 2007], where the sparse representations are obtained via a multi-scale dictionary learned using an example-based approach. The above de-noising algorithm has also been extended for color image restoration, de-mosaicing and in-painting [Mairal et al., 2008]. Finally, an over-complete dictionary design method that combines both the representation and the class discrimination power has been proposed in [Zhang & Li, 2010] for face recognition.
3. **Methods looking for "interesting" projections leading to learning projection matrices:** For instance, Linear Discriminant Projections (LDP) has been proposed as a useful tool for large-scale image recognition and retrieval [Cai et al., 2011]. Hashing has also been proven a computationally attractive technique, which allows somebody to efficiently approximate kernels for very high-dimensional settings by means of sparse projection into a lower dimensional space. For instance, in [Shi et al., 2009], hashing has been implemented for handling thousands of classes on large amounts of data and features. In the same fashion, specialized hash functions with unbiased inner-products that are directly applicable to a large variety of kernel methods have also been introduced. Exponential tail bounds that help explain why hash feature vectors have repeatedly led to strong empirical results are provided in [Weinberger et al., 2009]. The authors demonstrate that the interference between independently hashed subspaces is negligible with high probability, which allows for large-scale multi-task learning in a very compressed space. Finally, advances have been made in manifold learning through the development of adaptive techniques that address the selection of the neighborhood size as well as the local geometric structure of the manifold [Zhang et al., 2012].

Recently, **Compressed Sensing (CS)**, belonging to the third category, has emerged as a powerful mathematical framework that offers a reversible scheme for compressing and reconstructing a DO. The key difference from the rest of methods is that the projected matrices involved are learned using random (e.g. Gaussian) distributions, while other methods require intensive calculations in the training phase. The ability to use random, instead of complicated and difficult to generate, projection matrices for compressing a DO, offers a huge advantage in the context of big data and specifically with respect to velocity and memory related challenges, e.g. real-time processing, live streaming, etc. The use of CS has already been proven beneficial for a multitude of applications such as Image Processing (Annotation, De-noising, Restoration, etc.), Data Compression, Data acquisition, Inverse Problems, Biology, Compressive Radar, Analog-to-Information Converters and Compressive Imaging [Baraniuk, 2007].

3.2.2. Compressed Sensing Theory: Background

CS addresses the problem of compressing a high-dimensional signal using as few bits of information as possible. However, in this process, we need the appropriate guarantees that adequate information is encoded by the compressed data, and this premise is actually reflected to our ability to perfectly reconstruct the original signal from its compressed version. In the context of CS, such guarantees are provided through the assumption of the sparsity of the original data, where by sparsity we mean the extent to which data contain zero values.

Strictly speaking, given an *n*-dimensional signal **x** containing *k* non-zero coefficients, the aim is to project it onto a compressed *m*-dimensional signal **y**, where *m*<<*n*, using a projection matrix **D**, via **y = Dx**. As m denotes the number of reduced dimensions, and, thus, expresses the compression capacity, it is clear that we want it to be as small as possible. Towards this end, it is of great importance to appropriately design the projection matrix **D**, so that it enables subsequent reconstruction of the original signal. The great leap of CS, is proving that under the condition *m = O(k log(n/k))*, a matrix **D** whose entries have been randomly generated, e.g. using a Gaussian distribution, can serve as the appropriate projection matrix [Davenport et al., 2011]. From the previous condition, the smaller *k* is, the smaller *m* becomes, which can be restated as, the more sparse the original signal is, the more heavily this signal can be compressed. A block diagram illustrating the vector compression-reconstruction process is provided in Fig. 3-1.

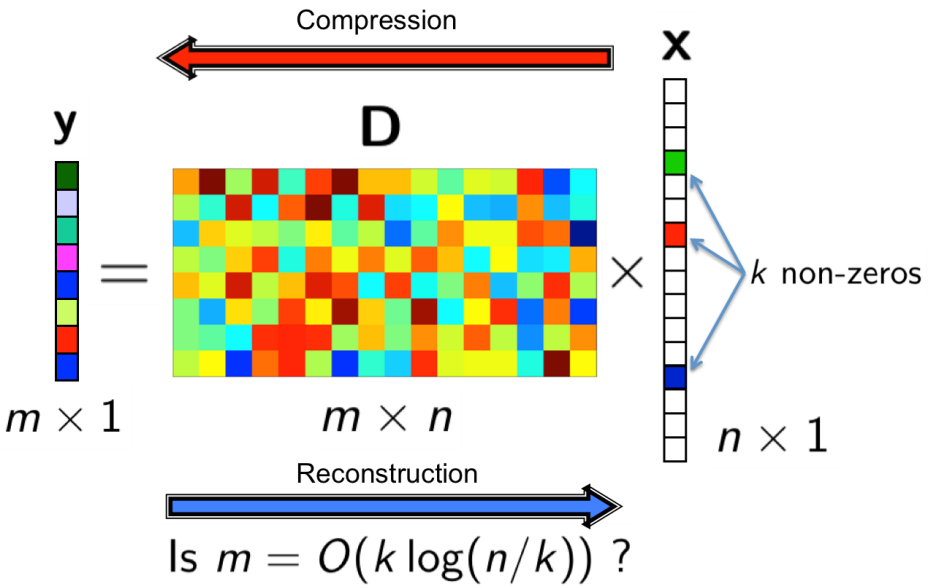


Fig. 3-1. Block diagram of CS.

Despite the impressive finding of CS that random projections qualify as the optimal choice in signal compression, the assumption of the data sparsity is very strict, and in reality sparse signals are rarely met in practice. Fortunately, the sparsity assumption can be replaced by the more relaxed assumption of compressibility allowing for almost lossless recovery of the original vector [Baraniuk et al., 2011; Candes & Walkin, 2008]. Compressibility measures the extent to which a vector can be approximated by a properly sparse signal and in fact, expresses the extent to which a vector contains coefficients close to zero. Having recognized the importance of sparsity/compressibility in compression-reconstruction using random projections, in the following section we review some indicative related work and propose a novel framework for measuring the sparsity of an arbitrary signal. From now onwards, for simplicity, when we use the term sparsity, we will refer to either sparsity or compressibility.

3.2.3. Calculating and Exploiting Data Sparsity

Data sparsity has been celebrated in signal compression as a solution to problems previously unsolvable. Roughly speaking, sparsity measures the extent to which the information of a vector is distributed to its coefficients. More specifically, for highly sparse vectors the information is concentrated to a small portion of coefficients, while for non-sparse ones the information is uniformly distributed across the coefficients. In this context, sparsity is a desirable property as it allows for succinct representations of large pieces of information.

There are many paradigms stemming from diverse research domains advocating the importance of sparsity. Clearly from the previous section, Compressed Sensing constitutes a vivid example where the role of sparsity has been demonstrated in the process of compressing and reconstructing a signal. More specifically, in the CS context, through the introduction of the Null Space Property (NSP) [Baraniuk et al., 2011] and the Restricted Isometry Property (RIP) [Candes & Walkin, 2008], it has been proven that the assumption of data sparsity allows for the perfect reconstruction of a signal that has been compressed using only a few random measurements of the original sparse signal. Towards this end, a variety of optimization algorithms incorporating the notion of sparsity has been proposed for reconstructing a compressed signal [Candes & Tao, 2007; Meinshausen & Yu, 2009].

Apart from the aforementioned applications, the notion of sparsity has also been incorporated in already existing methods in various fields. For instance, Bayesian methods providing sparse solutions to regression and classification problems have attracted a renewed interest [Tipping, 2001]. Moreover, in Support Vector Machines (SVM) optimal guarantees on the sparsity of the support vector set, encoding the boundary between two classes, have also been investigated [Cotter et al., 2013]. Sparsity appears to play a key role in boosting techniques as well, leading to sparse combinations of a number of weak classifiers [Xi et al., 2009]. Additionally, in an unsupervised configuration, Sparse Principal Component Analysis (S-PCA) has been introduced as a framework, which trades off redundancy minimization for sparsity maximization in the basis signals [Chennubhotla & Jepson, 2001].

Given the importance of sparsity in compression-reconstruction process, it is clear that a method to calculate or estimate the sparsity of a vector is required. A variety of methods have already been proposed in the bibliography for measuring the inherent sparsity of a DO. The most straightforward way to measure it is by using the l_0-norm. Although the l_0-norm has led to impressive theoretical results in sparse representation, in practice it suffers from several disadvantages [Karvanen & Cichocki, 2003]. For overcoming such disadvantages, approximations of the l_0-norm have been proposed in sparsity optimization problems in the presence of noise [Fuchs, 2005, Donoho et al., 2006]. Thresholding techniques have also been employed [Rath et al., 2008], however the selection of a reasonable threshold may prove to be problematic.

Due to its disadvantages, the l_0-norm has often been replaced by the l_1-norm, which offers a plausible alternative measure surpassing some of the shortcomings accompanying l_0-norm [Candes & Tao, 2005, Donoho & Tsaig, 2008]. Towards this direction, [Candes & Tao, 2005] comprises a milestone work, where the authors prove that the classical error correcting problem, under certain conditions can be translated into a l_1-optimization problem. The latter can be trivially solved in a linear programming configuration using existing dedicated methods. In a similar vein, in [Donoho & Tsaig, 2008] the authors employ the Homotopy method to solve an underdetermined system of linear equations through an l_1-minimization problem. In [Candes, 2008], the authors propose a methodology for sparse signal recovery that often outperforms the l_1-minimization problem by reducing the number of measurements required for perfect reconstruction of the compressed signal. The problem is decomposed into a sequence of l_1-minimization sub-problems, where the weights are updated at each iteration based on the previous solution.

Norms l_p of higher order have also been used for measuring signal sparsity [Karvanen & Cichocki, 2003]. Moreover, combinations of l_p norms have been proposed as well. The so called Hoyer sparsity measure based on the relationship between the l_1 and the l_2 norm has been presented in [Hoyer, 2004] in a sparsity constrained Non-negative Matrix Factorization (NMF) setting for finding linear representations of non-negative data. Apart from the l_p norms, other mathematical functions have also been used for measuring vector sparsity. For instance, kurtosis has been proposed as a metric for data following a unimodal and symmetric distribution form [Olshausen & Field, 2004]. In the same vein, in [Karvanen & Cichocki, 2003], the authors suggest the adoption of *tanh* functions as an approximate solution of l_p norms. Furthermore, they introduce a metric based on order statistics. In

contrast to l_p norms and similarly to our work, the functionality of the above methods is based on the distribution form of the signal coefficients rather than the magnitudes of the latter. A main drawback though is that they can only handle signals whose coefficients contain a unique dominant mode at zero, and thus must be avoided when dealing with data containing multiple modes, which constrains their scope of applications [Karvanen & Cichocki, 2003].

The connection between sparsity and entropy has been clearly demonstrated in [Pastor et al., 2013]. Entropy expresses the complexity of a signal, while sparsity expresses its compressibility under appropriate basis. Along these lines, the authors argue that both sparsity and entropy should follow similar intuitive criteria. Towards this end, they propose a novel sparsity and a novel entropy metric that satisfy such criteria. The functionality of these metrics is based on the calculation of the similarity of a signal to the theoretically totally non-sparse using the inner product between them. Relying on the above connection, entropy diversity metrics can also be used to measure sparsity [Rao & Kreutz-Delgado, 1999], [Kreutz-Delgado & Rao, 1998]. For instance, the Shannon and the Gaussian entropy presented in [Rao & Kreutz-Delgado, 1999] constitute plausible measures of sparsity, which incorporated in sparsity minimisation problems may lead to sparse solutions to the best basis selection problem [Kreutz-Delgado & Rao, 1998]. Finally, among the most prevalent sparsity metrics, the Gini Index (GI) offers a state-of-the-art solution, which has led to impressive results in recovering randomly projected signals [Gini, 1921, Zonoobi et al., 2011].

The majority of the above sparsity metrics have been motivated by intuitive incentives and their validity often relies merely on subjective criteria. For surpassing this limitation, in [Hurley et al., 2009, Pastor et al., 2013], a number of desirable objective criteria that a sparsity metric must satisfy has been proposed. These evaluation criteria provide a degree of credibility to a sparsity metric enabling the comparison between different metrics. [Hurley & Rickard, 2009], summarizes which of the sparsity criteria are satisfied by each of the sparsity metrics proposed in the literature. the mathematically rigorous definitions of the above-mentioned sparsity criteria, acquired from the literature, along with some intuitive interpretations are provided in the Appendix.

Generalized Differential Sparsity and its Exploitation in Signal Reconstruction

In this deliverable, motivated by the need to reconstruct sparse signals, which have been heavily compressed via CS, we suggest using the approach presented in [Zonoobi et al., 2011], which, combined with GI, has returned excellent results. In this context, the reconstruction is performed by incorporating a sparsity metric (e.g. GI) into a stochastic approximation method that solves a dedicated sparsity maximization problem. More specifically, given a compressed signal and based on the prior assumption that the original signal before compression was sparse, the idea is to find in the original space the signal with highest sparsity that gives the smallest reconstruction error. In this process, obviously the selection of sparsity metric plays a crucial role. For this purpose, we propose a novel **Generalized Differential Sparsity (GDS)** framework for generating novel sparsity metrics and we prove that this framework, incorporated in the previous reconstruction formulation, performs better than the GI and thereby the state-of-the-art.

For the remainder of this document, we will borrow the term vector from linear algebra in order to refer to a signal. So, let $\mathbf{c} = [c_1, ..., c_N] \in R^N$ be a vector whose sparsity we would like to measure. A sparsity metric is a function $S : R^N \rightarrow R$, which given an N-length real vector returns a real number that comprises an estimation of its sparsity. Note that the term metric used in this deliverable has not the typical meaning of metric, which obeys a number of mathematical properties. Instead, it means a function with the ability to measure. From now onwards, we implicitly assume that sparsity is measured using the magnitudes of the coefficients and not their algebraic values, i.e., the coefficient signs can be neglected. Therefore, for simplicity, we can assume that $c_i \geq 0$, $\forall i \in \{1, 2, ..., N\}$.

The central idea of sparsity is based on calculating how “small” the coefficients of a vector are. As a consequence, most of the already existing sparsity metrics use the magnitudes of the coefficients to

encode vector sparsity, defying the differences among these coefficients. However, it is clear that, how small or large a coefficient is, depends on what reference value it is compared with. Relying on this claim, we propose a sparsity metric that takes into account the relative differences between every pair of coefficients, thus comparing coefficients with each other. In this way, they achieve to measure the sparsity of a vector by examining the extent to which the energy is distributed to the coefficients. The following example illustrates why the relativity among coefficients may prove crucial.

Considering **v** = [0.001, 0.001, 0.001, 0.001, 0.001], two questions naturally emerge. How close are the coefficients of **v** to zero and thereby how sparse is **v**? The answer to the first question is that they are all equally close to zero, regardless of their actual distance from it. In this sense, all coefficients contain equal percentage of the vector energy, which means that all of them are equally important in the representation of the vector and, hence, none of them should be discarded. As a consequence, the answer to the second question is clearly that **v** is totally non-sparse, although this may look counter-intuitive at a first glance. Based on the intuition arising from the previous example, we give the definition of our GDS metric.

Definition of GDS:

The GDS of order p ($p \geq 1$) of a vector $\mathbf{c} \in \mathrm{R}^N$ is defined as:

$$S_p(\mathbf{c}) = \frac{1}{N\sum_{i=1}^{N} c_i^p} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (c_j - c_i)^p$$

where the coefficients have been sorted in ascending order so that: $c_1 \leq c_2 \leq \cdots \leq c_N$.

For different values the parameter p produces different GDS metric-instances. From the definition of GDS, we can easily prove the following Theorem, which imposes lower and upper bounds to the possible values of GDS regardless of the order p. The proof is given in the appendix.

Theorem 1

$$0 \leq S_p(\mathbf{c}) \leq 1 - \frac{1}{N}, \ \ \forall \mathbf{c} \in \mathbb{R}^N, \ \ \forall p > 1.$$

The following two theorems show that the order p of GDS determines the tendency of the corresponding GDS metric-instance to qualify an arbitrary signal as sparse. More specifically, as p increases, it is more difficult to qualify vectors as sparse by GDS. This proves to be a great advantage, since it offers GDS the flexibility to adjust to certain requirements arising from the nature of the data. The proofs of the theorems are again appended to the end of this deliverable.

Theorem 2

Given a vector $\mathbf{c} = [c_1,...,c_N]$, if $p > q$, then $S_p(\mathbf{c}) < S_q(\mathbf{c})$.

Theorem 3

For an arbitrary vector $\mathbf{c}$, we have that $\lim_{p \to +\infty} S_p(\mathbf{c}) = 0$.

The findings of Theorems 2 and 3 actually show that the order p determines the strictness of GDS in qualifying an arbitrary vector as sparse. This feature offers the appropriate granularity to GDS and allows it to adjust to certain circumstances stemming from the nature of the data and the specific problem to be solved. For instance, it is anticipated that for data containing few zeros, which are supposed to be inherently non-sparse, a large order might be needed to discriminate among different levels of sparsity. On the contrary, for data with plenty of zeros, smaller orders, i.e., less strict metrics might prove to be optimal. Therefore, finding and adopting the appropriate GDS metric to a sparsity maximization problem might lead to improved reconstruction results.

Closing this section, it is worth mentioning that GDS of first order is equivalent to GI. The proof can be found in the Appendix. The encapsulation of GI within GDS emphasises the power of the latter as a generalized framework for unifying already existing SotA sparsity metrics apart from generating novel ones.

A Computationally more Efficient GDS Formula for High-Dimensional Data

When the number of vector dimensions is large, the formula used for the definition of the GDS metric is difficult to compute. A more tractable and computationally efficient formula for GDS of order *p*, with *p* integer, is presented in this section. Moreover, a computational analysis is also provided in order to compare the two formulas. Due to computational reasons, the alternative formula is different for even and odd values of *p*. For this purpose, the two cases are separately presented. The rigorous proofs of the derivation of these formulas from the original one are provided in the Appendix.

Even formula:

$$S_{2k}(\mathbf{c}) = 1 + \frac{1}{N\|\mathbf{c}\|_{2k}^{2k}}\left[\sum_{\omega=1}^{k-1}(-1)^{\omega}\binom{2k}{\omega}\|\mathbf{c}\|_{\omega}^{\omega}\|\mathbf{c}\|_{2k-\omega}^{2k-\omega} + \frac{(-1)^k}{2}\binom{2k}{k}\|\mathbf{c}\|_{k}^{2k}\right]$$

Odd formula:

$$S_{2k+1}(\mathbf{c}) = \frac{1}{N\|\mathbf{c}\|_{2k+1}^{2k+1}}\gamma\,,$$

where

$$\gamma = \sum_{\omega=0}^{k}(-1)^{\omega}\binom{2k+1}{\omega}\sum_{i=1}^{N}\left(c_i^{2k+1-\omega}f_{\omega}(i) - c_i^{\omega}f_{2k+1-\omega}(i)\right)$$

and

$$f_{\omega}(i) = \sum_{j=1}^{i} c_j^{\omega}.$$

Computational Analysis:

It can be easily proven that the original formula requires $(p-1)\,N\,(N+1)/2$ multiplications and $N^2 - 2$ additions, which is in total on the order of $O(N^2p)$. This computational load is for many practical reasons inefficient when *N* is large. The corresponding load when using the "even" formula consists of $k^2(2N+1) - kN - k$ multiplications and $2kN - k + 2$ additions. For both even and odd formulas, the computational complexity is in total $O(k^2N)$, which is clearly more efficient when approximately $N > p/4$. However, in the opposite case, the original formula is more tractable.

Indicatively, for a 10000-dimensional vector, the even formula needs around 5000 times less calculations than the original one for $p = 2$ and 100 times less calculations for $p = 100$. The corresponding numbers for the odd formula are 1700 and 50 for $p = 1$ and $p = 99$, respectively. Summarising the above, for practical reasons, the new formula proves to be very useful in both even and odd cases. However, notice how the situation is reversed when *N* is very small. For example, for $N = 10$ and $p = 100$, the even and odd formulas need respectively around 10 and 20 times more calculations than the original formula. Therefore, both the original and the alternative formulas prove to be important and may be preferred according to the specific conditions.

Experimental Results

In this Section, using GDS, we investigate the reconstruction error of randomly projected sparse signals as a function of the order of GDS, the sparsity of the original data and the number of dimensions of the reduced data. More specifically, given an original vector $\mathbf{x_0} \in R^N$ and an $M \times N$ projection matrix A with $M < N$, $\mathbf{x_0}$ is projected to $\mathbf{y} = \mathbf{Ax_0} \in R^M$, and subsequently reconstructed to the initial space. In our work, the entries of **A** are generated using i.i.d random variables of a zero mean and unit standard deviation Gaussian distribution. The reconstruction of $\mathbf{x_0}$ from **y** is accomplished by solving the following constrained optimisation problem:

$$argmax\ S_p(\mathbf{x}) \text{ subject to } \mathbf{Ax} = \mathbf{y} \qquad (1)$$

where **x** is the estimate of the original vector and $S_p(\mathbf{x})$ is the sparsity of **x**. Essentially, having the prior information that the original signal before compression was sparse, the aim of Eq. 1 is to find the sparsest solution in the pool of feasible solutions satisfying the above constraint. For solving this problem, we have employed the iterative Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm [Spall, 1999]. Actually, for imposing the above constraints, we adopted the implementation presented in [Zonoobi et al., 2011], which combined with the GI has shown impressive performance in signal reconstruction. In our case, the parameters involved in SPSA have been selected based on some previous research results [Zonoobi et al., 2011; Sadegh & Spall, 1998].

There are two main reasons why we opted to use SPSA. First, SPSA does not make direct reference to the gradient of the objective function; instead, it approximates the gradient using only two calculations of the objective function per iteration, regardless of the signal dimensionality. This feature renders SPSA computationally very efficient in high-dimensional problems. Second, it has been proven that SPSA under general conditions converges to the global optima [Maryak & Chin, 2001].

In this experiment, we generated a random vector $\mathbf{x_0}$ of size $N = 100$ with K non-zero coefficients. We varied K in the range between 10% and 60% and the non-zero coefficients were generated using four different distributions: binomial, uniform, normal and exponential. It is worth noting that K and sparsity are inverse quantities, i.e. the smaller K is the sparser the vector is. For several values of the order p in the range between 1 and 10, we exhaustively varied the number M of the projected dimensions of **y** from 1 to 99 and we reconstructed $\mathbf{x_0}$ by employing SPSA/GDS using Eq. 1. Finally, for each setting, we calculated the Mean Square Error (MSE) between the recovered and the original vector. For ensuring statistical significance, we ran the whole above approach 100 times and we calculated the average MSE for each triple of values of K, M and p.

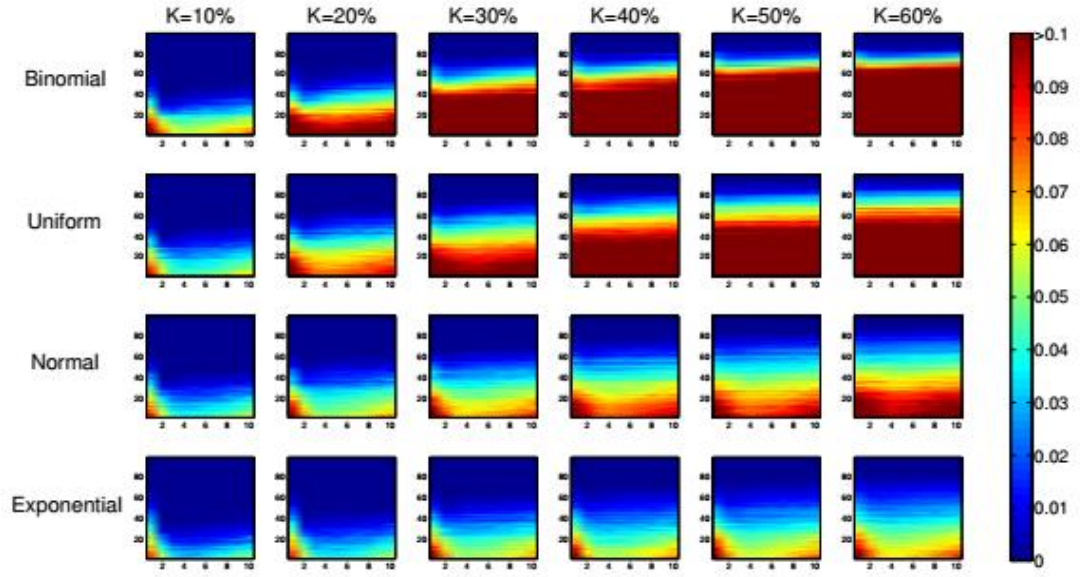


Fig. 3-2. Reconstruction error using various values for the order p of GDS and the number M of the reduced dimensions, for different values of the number K of non-zero coefficients of the original data. Horizontal axis: p, vertical axis: M, Colorbar: MSE.

The reconstruction errors that we obtained using the above settings are pooled in Fig. 3-2. The four rows correspond to the binomial, normal, uniform and exponential data, respectively. The figures of each row correspond to different values of K. For each figure, the horizontal axis depicts the order p

of GDS and the vertical axis the number *M* of projected dimensions. The MSE that corresponds to every pair (*p*,*M*) is indicated by the colorbar on the right.

Performing a row-wise (i.e., sparsity oriented) comparison in Fig. 3-2, it is interesting to observe that regardless of the data type, the greater *K* is, i.e., the less the sparsity of the original vector is, the larger the MSE becomes in general. This behaviour clearly verifies the importance of sparsity in signal reconstruction. Similarly, performing a column-wise (i.e., data type oriented) comparison, it is clear that regardless of *K*, the reconstruction error decreases as we move from top (binomial data) to bottom (exponential data). This can be attributed to the fact that for a specific *K*, although in all four cases we use equal number of non-zeros, in fact GDS tends to consider more sparse those vectors whose sorted absolute coefficients have larger differences. Fig. 3-3 illustrates the general form of an arbitrary vector generated using either of the above four distributions in the case where *K* = 40%. The first order GDS (i.e., GI) sparsities of these prototypic vectors are approximately 0.60, 0.73, 0.76 and 0.79, respectively. Apparently from the above, in terms of GDS, exponential distribution gives the sparsest vectors and thus the most eligible for reconstruction using the adopted methodology.

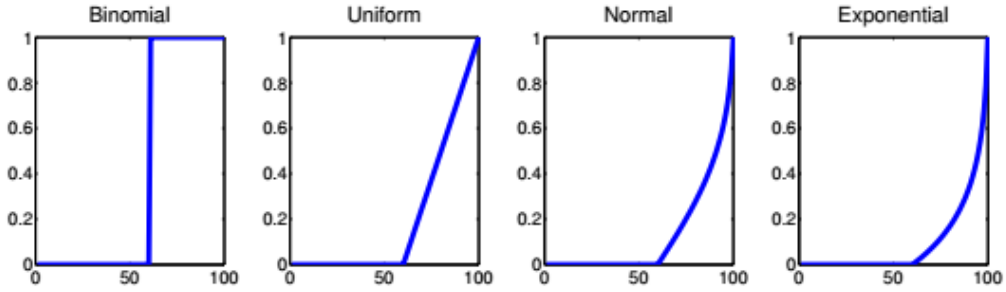


Fig. 3-3. General form of the absolute coefficients of signals generated using Binomial, Uniform, Normal and Exponential distributions, sorted in ascending mode.

Having a closer inspection at each subfigure, we observe that regardless of both *K* and data type, the reconstruction error has a similar form. More specifically, it is clear that in almost all cases - except for binomial and uniform with *K* > 40% - values of *p* in the range between 2 and 7 provide the best results, and this becomes more evident for small values of *M*. In this direction, our next concern was to quantify how the optimal *p* varies as a function of *M* and *K*, and Fig. 3-4 and Fig. 3-5 serve exactly this purpose. In Fig. 3-4, the horizontal axis depicts *M*, while the vertical axis contains the mean optimal *p*, as this has been calculated across the different values of *K*. It is clear in the figure that for small *M* (approximately <40), the best reconstruction is obtained by setting *p* between 4 and 6. Moreover, it is worth noticing that after a small reduction of the mean optimal *p* for intermediate values of *M* (i.e., in the interval 40-80), the superiority of high orders becomes more intensely evident for large values of *M*. However, for large *M*, the difference of reconstruction error among the several values of *p* becomes negligible as the MSE becomes almost zero. Finally, it is worth noticing that GI is never the optimal choice in reconstruction, justifying the use of higher orders and proving the superiority of GDS. In summary, the above findings explicitly demonstrate how GDS of high orders reduces the least number of projected dimensions required in order to perform almost perfect reconstruction of sparse signals.

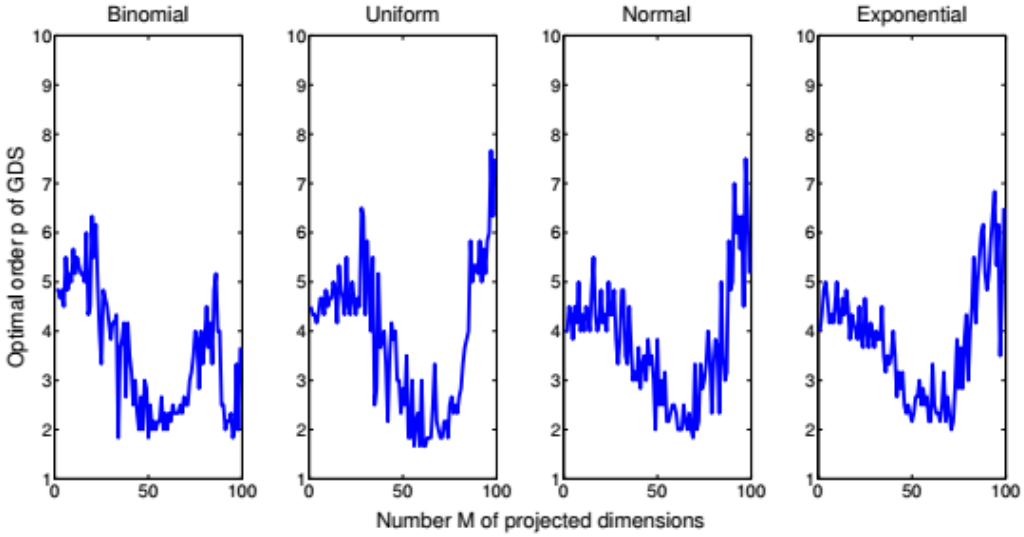


Fig. 3-4. Optimal order *p* of GDS as a function of the number *M* of projected dimensions.

Similar is the case when we investigate the optimal *p* as a function of the number *K* of non-zero coefficients. In Fig. 3-5, the horizontal axis depicts *K*, while in the vertical axis is the mean optimal *p*, as this has been calculated this time across the different values of *M*. Again, it is interesting to observe that GI never offers the best reconstruction performance. Instead, for every *K*, orders of GDS larger than 3 are needed. In particular, as the sparsity of the data decreases, larger values of the order *p* are required for better reconstructing a signal and on average $p = 4$ provides the best results. This outcome can be attributed to the strictness that *p* provides to GDS (cf. Theorems 2 and 3) and explicitly demonstrates how GDS loosens the bounds of the assumed sparsity of the original data offering more capacity in reconstructing lowly sparse signals.

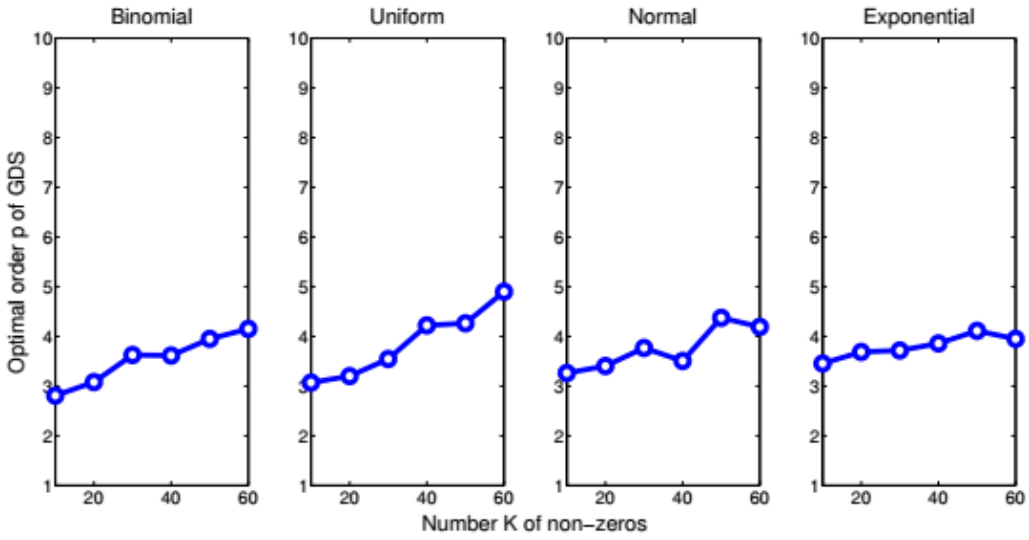


Fig. 3-5. Optimal order *p* of GDS as a function of the number *K* of non-zero coefficients.

3.3. Text-based Content Decomposition and Feature Extraction

Feature extraction and selection by statistical means is a standard content decomposition activity both on images and text. However, with text-based documents, we already have language in place for understanding semantic content, thereby one has to turn first to theories of word meaning from theoretical linguistics to model the evolution of document content over time. This will be briefly done below, starting with a new integrative model called the "field theory of semantic content", and next, interpreting its experimental results.

Field Theory of Semantic Content

It is universally accepted [Turney & Pantel, 2010] that the success of any vector space based classification method goes back to the contextual interpretation of word meaning [Wittgenstein, 1953; Firth, 1957], summed up by the distributional hypothesis which states that words typically used in the same context tend to have the same meaning [Harris, 1954]. By default, any DP method using vectors as mathematical objects to model semantic content falls back on this assumption of context-dependence. On the other hand, in CM, evolving systems of unevenly distributed physical content over spatiotemporal locations, like electromagnetism, or gravitation, are described by means of calculus, i.e. change is computed by differential equations and integral. Such systems are called fields in physics; therefore, we decided to model evolving semantic content on physical fields as a metaphor. This new idea was presented at conferences, and in publications [Wittek et al., 2014; Wittek et al., 2015a].

The major implication of using vector fields to represent categories of semantic content is that in vector space, only position vectors exist, pointing at located content. This means that such a model cannot address change. Fields, on the other hand, consist of two types of vectors, position and direction vectors, the latter type indicating directions of content modifications over time. A standard example is a map of plate tectonics in geology.

The field concept, and any tool based thereon, has important implications. Fields in physics contain energy and exert force. The Greek word energeia means work content, therefore the concept of semantic fields [Trier, 1934] implies that words with similar meanings and constituting regions of related content store the work equivalent of semantics. Further, the proposal has interpretations both from a CM perspective [Darányi & Wittek, 2012], and a QM angle [Wittek et al., 2014], the latter also including a distinction between Aristotelian dynamis vs. energeia, and is thereby suitable to detect new metadata as preservables. Prominently the Hamiltonian of evolving systems is anticipated as such a new preservable – this equation bridges the gap between classical and quantum systems, and describes the energy content of a system and its evolution. The Hamiltonian was shown to hold e.g. on social media by what the authors termed social mechanics [Lerman et al., 2011].

Experimental Results

Self-organizing maps (SOMs) are a widespread visualization tool that embeds high-dimensional data on a two-dimensional surface - typically a section of a plane or a torus - while preserving the local topological layout of the original data [Kohonen, 2001]. These maps provide a visual representation of groups of similar data instances, and they are also useful in analyzing the dynamics of evolving data collections, as updates are easily made. Emergent self-organizing maps contain a much larger number of target nodes for embedding, and thus capture the topology of the original space more accurately [Ultsch & Mörchen, 2005].

Training such a map is computationally demanding, but a great advantage of SOMs is that the computations are easy to parallelize. They have been accelerated on massively parallel graphics processing units [Luo et al., 2005]. Tools exist that scale to large data sets using cluster resources [Sul & Tovchigrechko, 2011], and also combining GPU-accelerated nodes in clusters [Wittek & Darányi, 2012]. These tools focus on batch processing. At the other end of the spectrum, interactive environments for training SOMs are often single-core implementations [Ultsch & Mörchen, 2005], not making full use of contemporary hardware.

For a practical analysis of context-dependent correlations, we have been developing a high-performance qualitative machine learning algorithm called **Somoclu**[4] which implements the above emergent self-organizing map approach. This tool is primarily meant for training extremely large emergent self-organizing maps on supercomputers, but it is also the fastest implementation running on a single node for exploratory data analysis. It can also be used for clustering based on self-organizing maps. The approach is generic and applies to any features represented by vectors, including image descriptors. We use this method for a variety of purposes: to analyze concept drifts, to demonstrate conceptual dynamics and a field theory of semantic change, and to uncover hidden correlations in sparse data collections. Further, Somoclu is already a mature software library with many users. It has a command-line interface, a C++ API, and wrappers in Python, R, and MATLAB.

Recall that a self-organizing map is a two-dimensional grid of artificial neurons. Each neuron is associated with a weight vector that matches the dimension of the training data. We take an instance of the training data, find the closest weight vector, and pull it closer to the data instance. We also pull the weight vectors of nearby neurons closer to the data instance, with decreasing weight as we get further from the best matching unit. We repeat this procedure with every training instance. This constitutes one training epoch. We repeat the same process in the second epoch, but with a smaller neighbourhood radius, and a lower learning rate when adjusting the weight vectors. Eventually, the neighbourhood function decreases to an extent that training might stop. The time needed to train an SOM grows linearly with the data set size, and it grows linearly with the number of neurons in the SOM. The resulting network reflects the local topology of the high-dimensional space [Kohonen, 2001]. Emergent self-organizing maps are a type of SOM that contain a much larger number of target nodes for embedding, and thus capture the topology of the original space more accurately [Ultsch & Mörchen, 2005]. Using a toroid map avoids edge effects.

Some nodes of the emergent self-organizing map correspond to one or more terms; these nodes or best matching units have a special role as they identify semantic content with one or more terms. The rest of the nodes act as an interpolation of the semantic field. Since the field is continuous in nature, we use toroid maps – a planar map would introduce an artificial discrete cut-off at the edges. To model an evolving field, we train a map on data from the first observation period, and continue training the same map as new instances enter from the next period. Hence the basic layout and the number of nodes do not change over time.

An important condition is that the total number of neurons, N_n, is much larger than the number of terms, N_t. For instance, given 12,000 terms, we expect to train a map with 60,000 neurons or more. The superfluous neurons will be the interpolation area for points not directly associated with a term (Fig. 3-6). Such maps are called emergent self-organizing maps [Ultsch & Mörchen, 2005], and they require highly parallel computational models [Wittek et al. 2015b].

[4] Somoclu is freely available under GNU GPL at https://goo.gl/AhTYCK

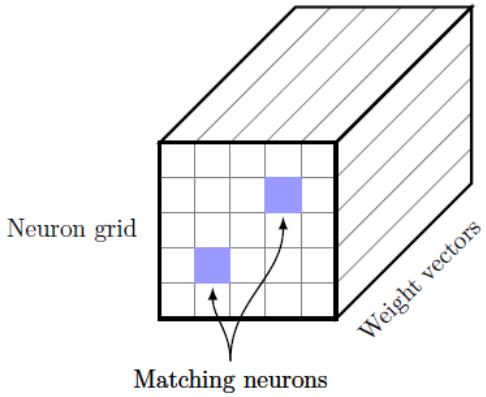


Fig. 3-6. A section of the two-dimensional surface of an emergent self-organizing map. At the end of the training, some neurons will correspond to index terms; these are shaded in the figure. Their corresponding weight vectors define the high-dimensional vector field at that point. Other neurons will interpolate the vector field between neurons that are associated with terms.

We used the Reuters-21,578 document collection of economic newswire as an example of text mining visualization [Lewis, 1999], starting with Lucene 3.6.2 to create an inverted index of the document collection. Terms were stemmed by the Porter stemmer, and we discarded those that occurred less than three times or were in the top ten per cent most frequent ones. Thus we had 12,347 index terms, lying in an approximately twenty-thousand dimensional space. We trained a toroid emergent self-organizing map of 336 × 205 dimensions. The initial learning rate was 1.0, which decreased linearly over ten epochs to 0.1. The initial radius for the neighbourhood was a hundred neurons, and it also decreased linearly to one. The neighbourhood function was a noncompact Gaussian.

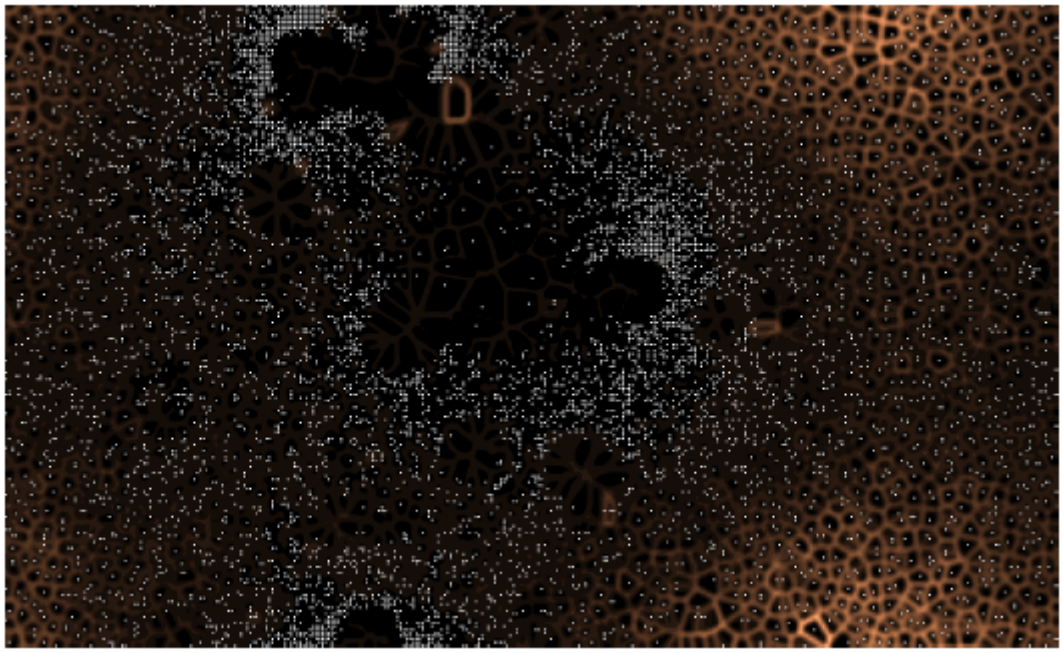


Fig. 3-7. The U-matrix of a toroid emergent self-organizing map after ten epochs of training on the feature space of sparse data. The individual dots are neurons with a weight vector that match a data instance. The other neurons reflect the distances in the original high-dimensional space.

We studied the U-matrix of the map. Visualizing this with the Databionic ESOM Tools , we plotted the global structure of the map in Fig. 3-7. The map clearly shows dense areas where index terms --

displayed as white dots, standing for best matching units (BMU) -- are close and form tight clusters. Other parts of the map are sparse, with large barriers separating index terms into individual semantic regions. Dark areas indicate inactive semantic content, strong boundaries refer to increasing distances between related terms in high-dimensional space with a tendency to split.

A first qualitative evaluation of the test results used the concept of semantic consistency, i.e. the relatedness of word meaning over a limited region in the map. E.g. Fig. 3-8 shows a tightly bundled group of terms. The gap between these words, based on the corpus, is small. The terms in this group, including ones that are not plotted in the figure, are: bongard, consign, ita, louisvill, occupi, reafffirm (with this misspelling), stabil, stabilis, strength, temporao, tight. Some are clearly related, for others, we need to look at the corpus for justification. The expression Bongard appears in two senses: the corrupt head of a bank, and as the name of a brokerage firm. ITA always refers to the International Tin Association, which was debating a pact extension at the time. Temporao is a kind of cocoa, firms trading it were listed on stock exchanges. Louisville as a location appeared frequently in the economic news typical in this test collection. The gaps are small between these words, which does not necessarily rule out the insertion of new words in the gaps, but based on the limited vocabulary a newswire, the lexical field represented by these expression appears to be covered.

Fig. 3-8. A cropped section of the U-matrix with best matching units and labels, showing a tight cluster. Some labels are not displayed as they overlapped with others.

Fig. 3-9. A cropped section of the U-matrix with best matching units and labels, showing large gaps between words.

On the other hand, large gaps are also interesting to look at. Take these examples (Fig. 3-9): energet, exploit, garrison, nato, petroleo. Apart from energet and garrison, these words are frequent, with over twenty occurrences in the collection each. The reason for their isolation is not because their corresponding term vectors do not contain entries. These words are related, but their best matching units were pulled in other directions, creating a tension in the lexical field. Over time, words labelling new content could be expected to emerge in such “red hot” topic zones where metaphoric fault lines separate cells containing terms displayed as white dots. Such fault lines manifest lexical gaps,

indicating content discontinuities in the observable field. In turn, such gaps vs. feature agglomerations are the kind of information we expect from statistical tools to help ontology construction.

3.4. Chapter Summary

The chapter presented our proposed feature extraction methodologies from visual and text-based content. Regarding the former, it has been clear that the pipeline approach for image representation leads to prohibitively massive volumes of data. Therefore, a compression methodology that enables a subsequent decompression of the data would be highly practical. As the previous pipeline leads to massive volumes of data, a compression-reconstruction methodology based on Compressed Sensing framework was proposed. Having recognized the crucial role of data sparsity in this process, a novel framework for measuring and exploiting sparsity was proposed. Via an experimental study, this new framework was tested on synthetic data and proved to enhance the functionality of the compression-reconstruction approach, extending the capacity of information that can be transferred from the original to the compressed data.

Regarding text-based content on the other hand, the chapter introduced a novel integrative model called the “field theory of semantic content”, which is based on established theories of word meaning from theoretical linguistics to model the evolution of document content over time. Along with presenting the basic notions behind this approach, we also presented a set of experimental results together with a discussion of the derived interpretations.

4. Semantic Concept Detection and Content Classification

In this chapter, we present our prototype methodologies for the extraction of semantic information from DOs, consisting of the detection of a number of semantic concepts from the content of a DO and the subsequent classification of this DO based on the previous information. Our approaches focus on the analysis of visual and text-based content, which are separately presented in the following subsections.

4.1. Analysis of Visual Content

It is generally accepted that high performance in visual classification schemes comes at the cost of increased computational complexity, formulating the known trade-off between effectiveness and efficiency. Examining this trade-off more closely, a typical pipeline for image classification consists of extracting features from a set of manually labelled images and using them within a classification scheme. Considering this pipeline there are two main scalability concerns.

First, the number of labelled training instances, which strongly affects the performance of the classifier, could hinder the scalability of an image classification pipeline. This significantly increases the computational cost of the classification scheme and the cost for manual annotation. Second, the dimensionality of the feature vectors could grow significantly, thus limiting the number of training instances that can be used (e.g. due to memory constraints). Consequently, it is imperative to find ways to minimize both the number of required training instances and the data dimensionality without sacrificing the performance of the classification models. Towards this goal, we present two methods in this section:

1. **SALIC (Social Active Learning for Image Classification)** is an approach that automatically gathers training data without requiring significant annotation efforts, while at the same time minimizing the number of the required training instances and increasing the performance of the classification models by utilizing a smart sampling approach.
2. **PCS (Product Compressive Sampling)** is a very fast method for dimensionality reduction, which yields similar results with the popular CS but only requires a small percentage of the time.

4.1.1. SALIC: Social Active Learning for Image Classification

It is commonly accepted that classification models become more robust when generated by high volumes of training data. However, the need for manually labelled training corpora creates a bottleneck for large-scale classification problems. In an effort to minimize the labelling effort, active learning [Cohn et al., 1994] trains an initial classifier with a very small set of labelled examples and expands the training set by selectively sampling new examples from a much larger set of unlabelled examples. These examples are selected based on their **informativeness**, i.e. how much they are expected to improve the classifier's performance. They are found in the uncertainty areas of the classifier and, in the typical case, are annotated upon request by an errorless oracle.

During the past decade, various active learning approaches have been developed using different sample selection strategies [Settles, 2010]. In [Chang et al., 2005], the selective sampling strategy is driven by the requirement to reduce the size of the version space as much as possible, by using an unlabelled sample that halves the version space. This way, the authors aim to cover the full unexplored space with the minimum number of queries. In [Hoi & Lyu, 2005], the authors attempt to

attack the insufficient training data problem by initially employing the semi-supervised approach until they collect a sufficient number of reliable training samples. They then deploy active learning for optimizing the sample selection process. In [Freytag et al., 2014], the authors propose a method to measure the expected change of the model outputs and utilize this measure as the informativeness of the new samples. In [Ebert et al., 2012], the authors analyse different sampling criteria and propose a method that adapts the sampling strategy during training by employing reinforcement learning. In [Li et al., 2013], the authors combine the typical uncertainty measure with an information density measure in order to define the **critical instances** to be labelled by the oracle.

On the other hand, the widespread use of online social networks has made available large amounts of user-tagged images that can be freely obtained and offer more information than their mere visual content (e.g. tags). If we could leverage these tags to become indicators of the actual content of the images, we could potentially remove the need for a human annotator and automate the active learning process. However, in this case, where the labels are leveraged from the freely available user tags, actively selecting new samples might seem superfluous, since there is no labelling cost to minimize. Indeed, we could simply just add all the images in the pool instead of actively selecting new ones. The question we should pose though is, *how many more images will we need to reach the same performance?* The computational overload of dealing with training sets several times larger might not seem important in the scale of a few concepts and a few thousand images. However, as we have witnessed in the past years, large scale image benchmarks gain an order of scale almost every year (from the ~10*k* images of the first Pascal-VOC competition in 2007 [Everingham et al., 2007], to the ~200*k* images of NUS-WIDE [Chua et al., 2009], to the 1*m* images of MIRFLICKR [Huiskes et al., 2010], to the 14*m* images of ImageNet [Deng et al., 2009a] and to the 100*m* images of Yahoo[5] nowadays). The situation is similar for the number of concepts, although the growth is not as steep (approx. from 20 concepts in 2007 to 20*k* nowadays). In this deliverable, we will prove that the ability to learn from social networks in an active manner is necessary to cope with the continuously growing requirements of large scale applications.

The idea of combining the benefits of active learning and knowledge generated from the crowds has recently become the focus of research. In this direction, Zhang et al. (2011) propose to use Flickr notes in the typical active learning framework (i.e. with a human oracle) with the purpose of obtaining a training dataset for object localization. In a similar endeavour, Vijayanarasimhan & Grauman (2014) introduce the concept of live learning, where they attempt to combine active learning with crowdsourced labelling. More specifically, rather than filling the pool of candidates with some canned dataset, the system itself gathers possibly relevant images via keyword search on Flickr. Then, it repeatedly surveys the data to identify the samples that are most uncertain according to the current model, and generates tasks on MTurk[6] to get the corresponding annotations.

However, although annotations originating from crowdsourcing services are closer to experts' annotations in terms of labelling accuracy [Nowak & Rüger, 2010], they cannot be considered either fully automated or free. In contrast, data originating from online social networks, although being noisier, can be used to support a fully automatic learning framework. In the direction of utilizing freely available web content, Golge & Duygulu (2014) propose a weakly supervised approach that collects data from web searches, applies clustering and outlier detection and trains a model from every cluster, assuming that it will represent a different characteristic of the concept. Web images are also used as the training set in [Li et al., 2014], where the authors treat the associated text as privileged information in a multiple instance learning scenario. Based on the assumption that search engines tend to return very relevant images in the first results of a query, Papadopoulou & Mezaris

[5] http://yahoolabs.tumblr.com/post/89783581601/one-hundred-million-creative-commons-flickr-images
[6] https://www.mturk.com/mturk/

(2015) create a set of queries, which is submitted to image search engines and the first images that are returned by each engine for each query are kept as positive examples.

However, tags obtained freely in this user-tagged images context cannot be considered as accurate labels. Combining the benefits of active learning dealing with user-tagged images, Li & Guo (2013) base their approach on the assumption that the tags of such images can reliably determine if an image does not include a concept, thus making social sites a reliable pool of negative examples. The selected negative samples are further sampled by a two-stage sampling strategy. First, a subset is randomly selected and then, the initial classifier is applied on the remaining negative samples. The examples that are most misclassified are considered as the most informative negatives and are finally selected to boost the classifier.

Furthermore, there have been several approaches that investigate active learning with noisy oracles; the objective is to model the expertise of each oracle, so that the most reliable for a specific instance can be selected. In this direction, Yan et al. (2011) consider active learning in a multiple oracles scenario, where the algorithm not only selects the informative samples but also the oracle to query labels from. In a similar scenario, Fang et al. (2014) propose the utilization of transfer learning in order to compute the reliability of each oracle by transferring knowledge from a different domain where labeled data can be abundant instead of depending on a large set of labelled images in the same domain. Rodrigues et al. (2014) extend the Gaussian Processes classification scheme to the multiple annotator scenario by treating the unobserved true labels as latent variables opting to estimate the different levels of expertise of the multiple annotators, obtaining in this way better estimates of the ground truth labels.

However, all these approaches assume that there is at least one oracle that holds the truth. This does not hold in the case of user-tagged images, since it is possible that the oracle (i.e. the web user) may never be absolutely confident of its decision. In such a setting, this adds a new factor (the oracle's *confidence* about the image's actual content) that should also be considered when selecting new samples. In contrast to other active learning approaches, **the novelty of our proposed SALIC framework lies in a sample selection strategy that maximizes not only the informativeness of the selected samples but also the oracle's confidence about their actual content**.

In order to achieve this goal, we formulate the selection process as a joint optimization problem that maximizes a function conditioned on the samples informativeness and the oracle's confidence. However, since this function cannot be estimated analytically, we propose a probabilistic approximation that leads to similar selection results. Towards quantifying this probability, we approximate the samples' informativeness by minimizing their distance from the separating hyperplane of the classification model, which is known to be an effective method for finding informative samples [Settles, 2010]. In order to measure the oracle's confidence, we propose the utilization of the popular bag of words approach [Joachims, 1998]. The reason for choosing this approach is its ability to decide about the content of an image based on a set of tags, thus capturing important contextual information and reducing the effect of erroneously provided, ambiguous and misleading tags. Joint maximization is then accomplished by ranking the samples based on the probability of a sample being selected given the two aforementioned quantities. This probability indicates the benefit that our system is expected to gain if the examined sample is selected and added to the training set.

PROBLEM FORMULATION

Let us consider the typical active learning scenario where we have a small set of manually labelled instances $L = \{l_1, l_2, ..., l_{NL}\}$, $N_L = |L|$ accompanied by their corresponding labels $Y = \{y_1, y_2, ..., y_{NL}\}$, $y_i \in D$, where D is the label space (the utilized notation is summarized in Table 4-1). In our case, we consider the binary classification problem (i.e. $D = \{+1, -1\}$), thus the probable labels are positive (i.e.

$d_1 = +1$) and negative (i.e. $d_2 = -1$), with respect to the concept that we are trying to learn. In addition to the labelled set, we have a large set of unlabelled instances $U = \{u_1, u_2, ..., u_{NU}\}$, $N_U = |U|$. Moreover, there is an oracle O providing labels accurately for the unlabelled set U on demand (i.e. $y_i = O(u_i)$). Initially, a baseline classifier H_0 is trained on the labelled set L. The objective of active learning is to establish a function $E(H_0, U)$, which defines the informativeness of each instance in the unlabelled set based on the previous classifier H_0. The instance $u^* \in U$ maximizing this function is selected to be added in the labelled set L along with its label y^*, which is provided by the oracle O (i.e. $y^* = O(u^*)$).

Table 4-1. Notation table.

Symbol	Definition
$L = \{l_1, l_2, ..., l_{NL}\}$	The set of labelled images.
$Y = \{y_1, y_2, ..., y_{NL}\}$	The labels of L.
$D = \{+1, -1\}$	The label space.
$U = \{u_1, u_2, ..., u_{NU}\}$	The set of unlabelled images.
$T = \{\mathbf{t_1}, \mathbf{t_2}, ..., \mathbf{t_{NU}}\}$	The tags of U, where $\mathbf{t_i}$ is the set of tags that correspond to image u_i.
$\mathbf{u_i}$	The visual descriptor of the image u_i.
$\mathbf{u}^{\mathbf{text}}_{\mathbf{i}}$	The textual descriptor of the image u_i.
$H_m = \{\mathbf{w_m}, b_m\}$	The classifier of iteration m ($\mathbf{w_m}$ is the normal vector to the SVM hyperplane and b_m is the bias term). H_0 is the baseline classifier, trained with the initially labelled data.
$E(H_m, u)$	The informativeness of sample u based on classifier H_m (i.e. the classifier at iteration m).
$O(u, t, d_k)$	The confidence of the oracle that the sample u belongs to the label d_k based on its tags t.
$S_i \in \{0,1\}$	The random variable (RV) modelling the event of selecting the image u_i.
$V_i \in [0,1]$	The RV modelling the probability that u_i is informative.
$T_i \in [0,1]$	The RV modelling the probability that u_i belongs to the examined concept $d_k \in D$.

*we use normal letters (e.g. u_i) to indicate individuals of some population and bold face letters (e.g. $\mathbf{u_i}$) to indicate vectors or sets of individuals of the same population

In our case, the oracle is substituted with the tags of the web users. Thus, the set of images U is not completely unlabelled but is associated with a set of tags $T = \{\mathbf{t_1}, \mathbf{t_2}, ..., \mathbf{t_{NU}}\}$, where $\mathbf{t_i}$ is the set of tags associated with image u_i. In this sense, the oracle has already answered for all the instances in U beforehand (i.e. the web users have already tagged the images). However, given the noisy nature of the user tags, we consider the oracle to be of questionable reliability and instead of providing an accurate label for a specific sample u^* as before, it provides its confidence for each sample u_i and label $d_k \in \{+1, -1\}$ (i.e. its confidence $O(u_i, \mathbf{t_i}, d_k)$ that the sample u_i is positive if $d_k = +1$ and negative if $d_k = -1$) based on the tags $\mathbf{t_i}$ of the image u_i. The objective in our case is not only to find the most informative sample, but also the sample for which the oracle is most confident for its label, so that we do not add falsely annotated samples in the training set. In order to do this, we have to jointly

maximize the informativeness of a sample $E(H_0,U)$ and the confidence of the oracle $O(U,T,d_k)$ for the label d_k. Thus we have to establish a new function G, which defines the benefit of selecting a sample u_i given its informativeness and the oracle's confidence. Maximizing this function, we can select the optimal sample u^* with the maximum informativeness and the higher confidence of the oracle for its label:

$$u* = arg \max_{u_i \in U} G(u_i | E(H_0, u_i), O(u_i, t_i, d_k)) \quad (1)$$

The instance u^* that maximizes this function for the examined label d_k is added to the training set as positive if $d_k = +1$ or negative if $d_k = -1$; the expectation is that the addition of such samples in the training set will allow for the maximum performance gain of the classifier. With this approach, positive and negative instances can be selected independently, by setting $d_k = +1$ for the positive and $d_k = -1$ for the negative in Eq. 2.

Considering that active learning is an iterative method, in the next iterations $m \geq 2$ the informativeness is defined as $E(H_m,u)$. For simplicity, in the following, we will show the methodology for the first iteration using the baseline classifier H_0, while the full iterative approach can be seen in Algorithm 1.

Active Learning with an Unreliable Oracle

Given that the function $G(u_i|E(H_0,u_i),O(u_i,\mathbf{t_i},d_k))$ cannot be analytically estimated, we choose to approximate it as a probability. For this reason, let us denote the following random variables (RV); S_i, the RV modelling the event of selecting the image u_i ($S_i \in \{0,1\}$), V_i, the RV modelling the probability that u_i is informative ($V_i \in [0,1]$), T_i, the RV modelling the probability that u_i belongs to the examined concept d_k ($T_i \in [0,1]$). Without loss of generality, we can assume that the function $G(u_i|E(H_0,u_i),O(u_i,\mathbf{t_i},d_k))$ is proportional to the probability of selecting an instance ($S_i = 1$) given its informativeness (V_i) and the confidence of the oracle (T_i):

$$G(u_i|E(H_0,u_i),O(u_i,\mathbf{t_i},d_k)) \sim P(S_i = 1|V_i,T_i) \quad (2)$$

Consequently, in order to find the optimal u^*, instead of the function G in Eq. 2, we can maximize its proportional probability function $P(S_i = 1|V_i,T_i)$. In order to calculate this probability, we make the reasonable assumption that the probability of an image being informative is conditionally independent from the probability that this image belongs to the examined concept (i.e. V_i and T_i are conditionally independent). Using Bayes rule and based on our assumption that V_i and T_i are independent we can express the probability $P(S|V,T)$ as follows (from now on the subscripts of S, V and T will be omitted):

$$P(S|V,T) = \frac{P(V,T|S)P(S)}{P(V,T)} = \\ = \frac{P(S|V)P(S|T)P(V)P(T)}{P(V,T)P(S)} \quad (3)$$

In calculating $P(S|V,T)$ we may encounter two cases:

a) If **$P(S = 0|V,T) \neq 0$**:: In order to calculate the probability $P(S = 1|V,T)$ and eliminate the probabilities $P(V)$, $P(T)$ and $P(V,T)$, we divide the probability of selecting an image with the probability of not selecting it, following the methodology presented in [Kordumova et al., 2014].

$$\frac{P(S=1|V,T)}{P(S=0|V,T)} = \frac{\frac{P(S=1|V)P(S=1|T)P(V)P(T)}{P(V,T)P(S=1)}}{\frac{P(S=0|V)P(S=0|T)P(V)P(T)}{P(V,T)P(S=0)}}$$

Then by replacing $P(S = 0|V,T)$, $P(S = 0|V)$, $P(S = 0|T)$ and $P(S = 0)$ with their complements ($1 - P(S = 1|V,T)$, $1 - P(S = 1|V)$, $1 - P(S = 1|T)$ and $1 - P(S = 1)$ respectively), we get the following equation that computes $P(S = 1|V,T)$:

$$P(S=1|V,T) = \frac{P(S=1|V)P(S=1|T)}{P(S=1) - P(S=1)P(S=1|T)} \cdots$$
$$\frac{(1 - P(S=1))}{-P(S=1)P(S=1|V) + P(S=1|V)P(S=1|T)} \quad (4)$$

b) If **P(S = 0|V,T) = 0***::* Assuming that the probabilities $P(V)$ and $P(T)$ cannot be 0\ (i.e. there is always an a-priori probability that an image is informative and belongs to the examined concept respectively), from Eq. 3 we have that either $P(S = 0|V) = 0$ or $P(S = 0|T) = 0$ (i.e. $P(S = 1|V) = 1$ or $P(S = 1|T) = 1$). Note that, in this case, Eq. 4 also produces the same result (i.e. $P(S = 1|V,T) = 1 \Rightarrow P(S = 0|V,T) = 0$), so from now on we will use Eq. 4 in all cases.

Thus, we only need to estimate three probabilities: $P(S = 1)$, $P(S = 1|V)$ and $P(S = 1|T)$. The first one is set to 0.5 as the probability of selecting an image without any knowledge is the same with the probability of dismissing it. For the probability of selecting an image given its informativeness $P(S = 1|V)$, we will approximate it using the informativeness criterion E, which can be any criterion of the Active Learning theory. In our case, we will be using the popular criterion that considers the most informative samples to be the ones lying on the separating hyperplane, since Support Vector Machines (SVMs) will be used to train the classification models. For the probability of selecting an image given the oracle's confidence $P(S = 1|T)$, we will approximate it with a textual analysis algorithm that takes as input the tags of an image and provides as output the probability of an image being positive or negative with respect to a concept, as explained below. While any textual analysis algorithm can be used, in this work, we used the popular Bag-of-Words scheme (BoW) [Joachims, 1998] due to its ability to additionally consider the context of the tags.

Incorporating informativeness (P(S|V))

The probability $P(S = 1|V)$ can be approximated using the informativeness criterion $E(H_0,u)$, where H_0 is the baseline classifier. As mentioned above, SVMs were chosen as the classification scheme for this work. For the SVMs, a popular informativeness criterion dictates the selection of the samples closest to the separating hyperplane as the most beneficial samples [Chang et al., 2005]. While other criteria could be used (e.g. halving the version space by explicitly calculating it when a sample is added), they would require to train a huge number of SVM models in order to calculate the version space. On the other hand, the selected criterion can be calculated very efficiently for linear SVMs since it requires only a dot product to calculate the distance from the hyperplane, making it very attractive for large scale pool of candidates.

Initially, the baseline classifier H_0 is trained using the manually annotated set L as a linear SVM classifier (i.e. $H_0 = \{\mathbf{w}, b\}$, where w is the normal vector to the hyperplane and b the bias term). Then, for every candidate image in the pool of candidates $u_i \in U$, the distance from the hyperplane $V(H_0,u_i)$ is extracted by applying the SVM classifier ($\mathbf{u}_i$ here denotes the feature vector of the image u_i):

$$V(H_0,u_i) = \langle \mathbf{w}, \mathbf{u}_i \rangle + b \quad (5)$$

Based on [Chang et al., 2005], the samples with the minimum distance to the hyperplane are considered as the most informative ones while the samples that lie outside the margin area of the SVM model are not expected to have any impact on the classifier. Thus the function $E(H_0,u_i)$, and consequently the probability $P(S|V)$, should be maximized (i.e. $P(S|V) = 1$) for the samples lying at the hyperplane and minimized (i.e. $P(S|V) = 0$) for the samples that are outside the margin area.

Based on the above observation, we approximate the probability $P(S|V)$ with the following equation:

$$P(S|V) \sim \begin{cases} 1-|V| & \text{if } -1<V<1 \\ 0 & \text{else} \end{cases} \quad (6)$$

where $|V|$ is the absolute value of the quantity in Eq. 5.

Measuring oracle's confidence (P(S|T))

In order to measure the oracle's *confidence* $O(u_i, \mathbf{t_i}, d_k)$ that the image $u_i \in U$ belongs to class d_k (i.e. is positive if $d_k = +1$ or negative if $d_k = -1$), we incorporate the associated textual information that is provided in the form of tags. Opting to overcome the noisy nature of social tagging (i.e. lack of structure, ambiguity, redundancy, emotional tagging, etc), we propose the utilization of the popular bag-of-words scheme [Joachims, 1998], due to its ability to capture the context of the whole set of tags, instead of only the meaning of each tag independently (e.g. as in the case of tag-to-tag similarity based on WordNet [Fellbaum, 1998]).

The vocabulary is extracted from a large image dataset crawled from flickr. Initially the distinct tags of all images are gathered. The tags that are not included in WordNet are removed and the remaining tags compose the vocabulary. Then, in order to represent each image with a vector, a histogram is calculated by assigning the value 1 to the bins of the image tags in the vocabulary and 0 to the rest. Principal Component Analysis (PCA) [Jolliffe, 2002] is applied to reduce the high dimensionality of the initial vocabulary ($\sim 47k$ distinct tags) to $7k$, which was chosen so that 95% of the variance is kept.

Afterwards, a linear SVM model ($\mathbf{w}^{text}, b^{text}$) is trained using the tag histograms as the feature vectors. In order to do this, a training set of images that contain both tags and manual annotations is utilized. The tags are required in order to calculate the feature vectors and the manual annotations to provide the class labels for training the model. In the testing procedure, for every tagged image $u_i \in U$ the feature vector u^{text}_i is calculated as above and the SVM model is applied. This results in a value for each tagged image $T(u_i)$, which corresponds to the distance of u^{text}_i from the textual hyperplane:

$T(u_i) = \langle \mathbf{w}_{text}, \mathbf{u}_{text} \rangle + b_{text}$

Thus, the oracle's confidence $O(u_i, \mathbf{t_i}, d_k = +1)$ that the image u_i is positive, and consequently the probability $P(S|T)$ when selective positive examples, can be calculated using Platt's sigmoid [Platt, 1999]:

$$P(S|T) = \mathcal{O}(u_i, \boldsymbol{t_i}, d_k = +1) = \frac{1}{1+\exp(AT+B)} \quad (7)$$

The parameters A and B are learnt on the training set using cross validation. The probability of Eq. 6 is for selecting a positive image. In the case that we want to select negative images, the oracle's confidence $O(u_i, \mathbf{t_i}, d_k = -1)$ that the image u_i is negative, and consequently the probability $P(S|T)$ when selective negative examples, is set to be the complement of Eq. 6:

$P(S|T) = O(u_i, \mathbf{t_i}, d_k = -1) = 1 - O(u_i, \mathbf{t_i}, d_k = +1)$ (8)

Retraining

In each iteration, after ranking the images based on the probability $P(S = 1|V,T)$, the top b_+ positive and b_- negative examples are selected to enhance the training set. In the case where both positive and negative samples are selected (as in SALIC), a new classifier is trained using the union of the old data and the newly selected examples as training set. However, in the case where either only positive or only negative examples are selected (as in some baselines we compare with), the classes become imbalanced, which is a typical problem in the machine learning field. In order to cope with the class imbalance problem we also apply a model aggregation method [Li et al., 2013].

The complete algorithmic procedure of SALIC can be seen in Algorithm 1.

Algorithm 1 SALIC

Input:
labelled data L with their labels Y,
unlabelled data U,
an oracle O,
the number of iterations N,
the number of positive and negative instances selected in each iteration b+, b−.

Output:
a classifier H_N.

1: $H_0 = \{\mathbf{w_0}, b_0\} = train(L, Y)$
2: **for** m = 1 : N **do**
// Get Positive Instances
3: $d_k = +1$;
4: **for** j = 1 : b+ **do**
5: Choose $u^* = \underset{u \in U}{\text{argmax}}\, G(u|E(H_{m-1}, u), O(u, \mathbf{t}, d_k))$
G is calculated using equations 2, 4, 6 and 7
6: $\{L, Y\} \leftarrow \{L, Y\} \cup \{u^*, y^* = d_k\}$
7: $U \leftarrow U \setminus u^*$
8: **end for**
// Get Negative Instances
9: $d_k = -1$;
10: **for** j = 1 : b− **do**
11: Choose $u^* = \underset{u \in U}{\text{argmax}}\, G(u|E(H_{m-1}, u), O(u, \mathbf{t}, d_k))$
G is calculated using equations 2, 4, 6 and 8
12: $\{L, Y\} \leftarrow \{L, Y\} \cup \{u^*, y^* = d_k\}$
13: $U \leftarrow U \setminus u^*$
14: **end for**
15: $H_m = \{\mathbf{w_m}, b_m\} = train(L, Y)$
16: **end for**

EXPERIMENTS

Two datasets were employed for the purpose of our experiments:

- The imageCLEF dataset *IC* [Thomee & Popescu, 2012] consists of 25*k* manually labelled images and was split into two parts, IC_{train} and IC_{test} consisting of 15*k* train and 10*k* test images respectively. The dataset was annotated by a vocabulary of 94 concepts which belong to 19 general categories[7]. The images of this dataset originate from flickr and the tags were also provided. From this dataset we obtained the manually labelled dataset L and the evaluation set (IC_{test}).
- The MIRFLICKR-1M dataset *F* [Huiskes et al., 2010] consists of one million user tagged images harvested from flickr. The images of *F* were tagged with 862115 distinct tags of which 46937 were meaningful (included in WordNet). Given that the *IC* dataset is a subset of *F*, the images

[7] age, celestial, combustion, fauna, flora, gender, lighting, quality, quantity, relation, scape, sentiment, setting, style, time of day, transport, view, water, weather.

that are included in both sets were removed from *F*. In our experiments, this dataset constitutes the pool of user tagged images.

For the visual representation of the images, we have used two popular approaches in order to verify the effectiveness of SALIC in different conditions; one that results in very high dimensional features that are of medium performance and one for low dimensional features that have shown remarkable performance. Our objective is to examine whether this difference in the discrimination ability of the two feature spaces will lead to particular requirements with respect to the sample selection process.

First, for the high dimensional features we have used the approach that was shown to perform best in [Chatfield et al., 2011]. More specifically gray, SIFT features were extracted at densely selected key-points at four scales, using the vl-feat library [Vedaldi, 2007]. Principal component analysis was applied on the SIFT features, decreasing their dimensionality from 128 to 80. Then, Fisher vector encoding (256 GMM components) and spatial pyramids (1×1, 3×1, 2×2 regions) were applied, resulting in a 327680 – *dimensional* feature vector per image.

Second, the low dimensional features are extracted from Convolutional Neural Networks (CNNs), which have demonstrated remarkable performance in both image annotation and object detection [Krizhevsky et al., 2012]. The implementation and the pre-trained CNN models of [Chatfield et al., 2014] were used. More specifically, the models for extracting the lowest dimensional feature vectors were utilized (model CNN M 128), resulting in only 128 dimensions.

Linear SVM models were trained for both feature spaces using the LIBSVM library [Chang & Lin, 2011] and were evaluated by mean Average Precision (mAP). The implementation code along with the data to reproduce the results are available at [8].

Considering that the utilized datasets consist of images labelled for multiple concepts, in order to conform with the binary classification requirements of SALIC, we transformed the multi label scheme to binary using the one-vs-all approach. More specifically, for every concept, as positive samples were considered all images including this concept in their list of annotated labels, leaving all the rest of the images as negative samples. Thus, the same image could serve as a positive sample for more than one concept. In our experimental study, 100 positive and 100 negative images from the 15k training images of IC_{train} were randomly selected to train the baseline classifiers (from now on called the IC_{init} dataset). The same initial examples were used for all experiments to allow for a fair comparison. Then, in each iteration, 100 images were selected from the pool of candidates to enhance the training set. As shown by Li et al. (2013), such iterative processes are robust to parameter changes. For this reason, and in order to compare with [Li et al., 2013], we start with an initial training set consisting of 100 positive and 100 negative examples and always add 100 examples in each iteration (i.e. either 100 positive or 100 negative or 50 positive and 50 negative depending on the training set expansion strategy). Finally, 50 iterations were conducted in all cases, i.e. in total, 5k images were added for each independent binary classification problem (i.e. concept).

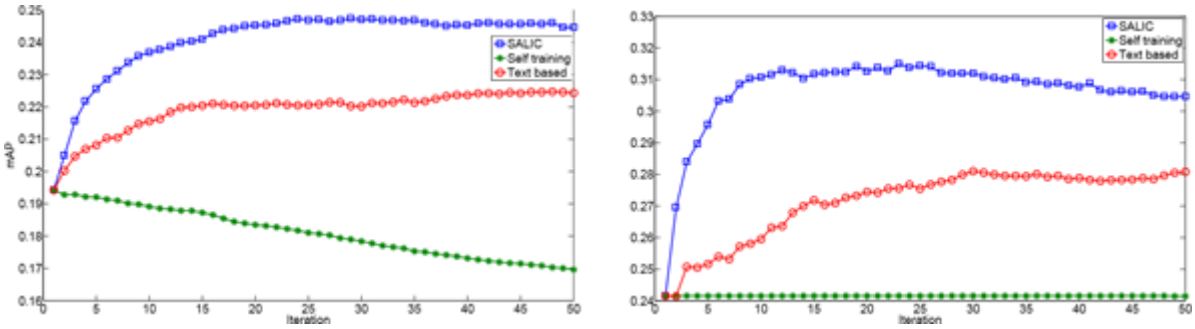
Comparing with Sample Selection Approaches

Quantitative evaluation of the proposed selective sampling approach: Here we compare the performance boost achieved by SALIC with two selective sampling baselines: a) self learning [Ng & Cardie, 2003], where the images that maximize the certainty of the SVM model (Eq. 5) trained on visual information are selected to expand the training set, and, b) text-based, where the images that maximize the oracle's *confidence* (Eq. 7 for positive and Eq. 8 for negative) are selected.

According to the results (see Fig. 4-1), self learning does not provide any benefit to the models; this can be explained by the fact that adding examples far away from the hyperplane does not add any information to the model. Moreover, the fact that SALIC outperforms both approaches indicates the

[8] http://mklab.iti.gr/research/salic

importance of incorporating the informativeness of new images in the selection strategy. With respect to the two utilized feature spaces, we can see that the results are very similar, with the only difference of self learning. In the case of the CNN features, the updated models do not gain any performance compared to the baseline ones. This can be attributed to the high discrimination ability of these features that forces self learning to select images far away from the hyperplane and thus not producing additional support vectors. On the other hand, in the case of the Fisher features, although the hyperplane of the self learning approach changes through the iterations, we can see that its performance deteriorates instead of improving. This can be attributed to the fact that the confidence of the oracle is not taken into account and thus the algorithm selects false positives/negatives to add in the training set.



(a) Fisher-based features.

(b) CNN-based features.

Fig. 4-1. Comparing with sample selection baselines.

Qualitative evaluation of the proposed selective sampling approach*:* Here we qualitatively compare SALIC with the typical self learning approach. In order to do so, we show visually the positive and negative samples that have been selected by the two approaches for concept *coast*. The CNN-based features are used for this visualization (Fig. 4-2). In each figure we show 20 randomly positive (or negative) examples from the initial training set and the top 20 positive and negative images that were chosen by each strategy. For the positive examples (Fig. 4-2a), we can see that the initial training set mostly consists of calm beaches depicting the sea, the sun and the sky, without many additional objects. As expected, the self learning strategy also selects very similar images to the training set, whereas the proposed approach selects images that depict rocky beaches or wavy seas with more objects (e.g. house, boat, lifeguard tower, surf boards and people). By adding images with larger content variety, the generalisation ability of the enhanced classifiers is maximized. Similarly for the negative examples (Fig. 4-2b), we can see that the proposed approach selects more relevant negatives (e.g. underwater images, blue buildings, skies with fields, etc.), whereas the self learning approach mostly selects indoor images that are completely irrelevant to the examined concept.

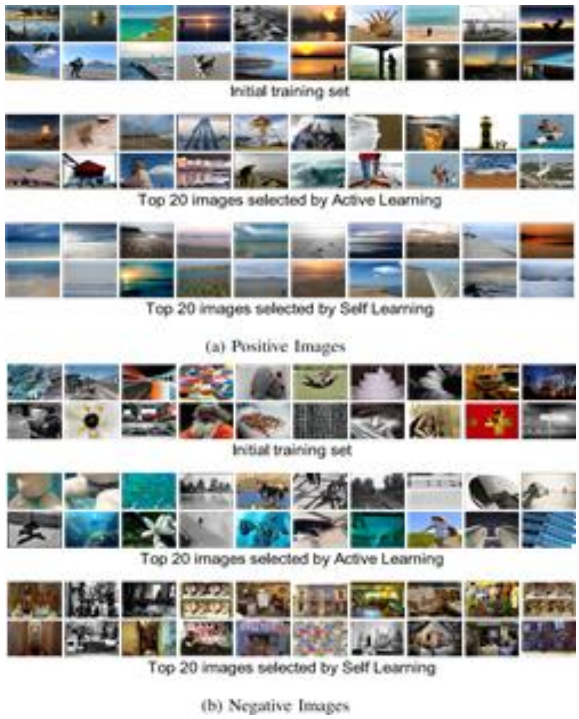
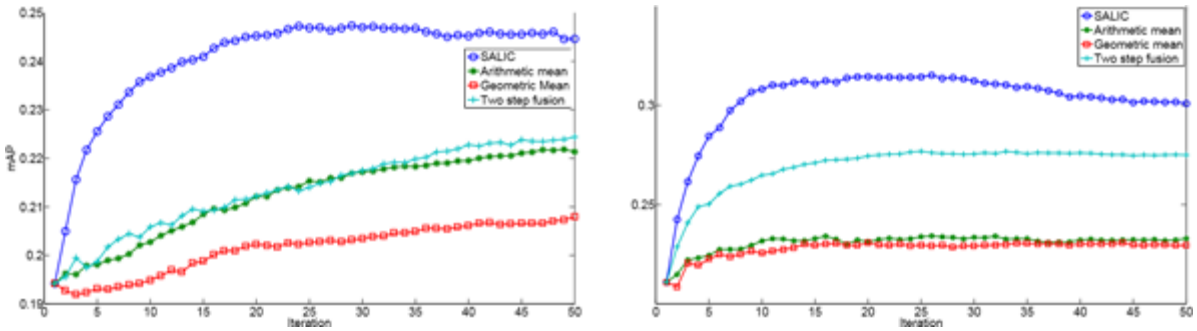


Fig. 4-2. Selected (a) positive and (b) negative examples, by active learning and self training for the concept coast.

Comparing with Fusion Approaches

In this section, in order to show the benefit of utilizing the proposed probabilistic approach presented in Section IV for fusing the informativeness of the samples $P(S|V)$ and the oracle's confidence $P(S|T)$, we compare it with three baseline fusion strategies: a) the arithmetic mean, where the two probabilities are combined into the selection probability $P(S|V,T)$ via the arithmetic mean, b) the geometric mean, where the two probabilities are combined into the selection probability $P(S|V,T)$ via the geometric mean and c) a two step approach that simulates the typical way that active learning is performed, while also keeping the classes balanced. In this case, in the first step, the tagged images in the pool of candidates are annotated as positive or negative based on the oracle's confidence $P(S|T)$.

In the second step, the most informative of the positive and negative images are selected (i.e. the images that maximize the probability $P(S|V)$). The results are shown in Fig. 4-3 both for Fisher- (Fig. 4-3a) and CNN- (Fig. 4-3b) based features. We can see that using the simplistic arithmetic and geometric mean approaches does not improve the initial models significantly, showing that the selected samples are either not informative or false positives/negatives. On the other hand, the two step approach yields a higher performance boost of the models, which is probably due to the strict filtering it applies on the tagged images so that false positives/negatives are not selected, while keeping the notion of informativeness in the selection process. On the contrary, SALIC greatly outperforms all the baseline fusion strategies by selecting the samples that will have a higher impact when added in the training set, showing the importance of maximizing the joint probability $P(S|V,T)$.

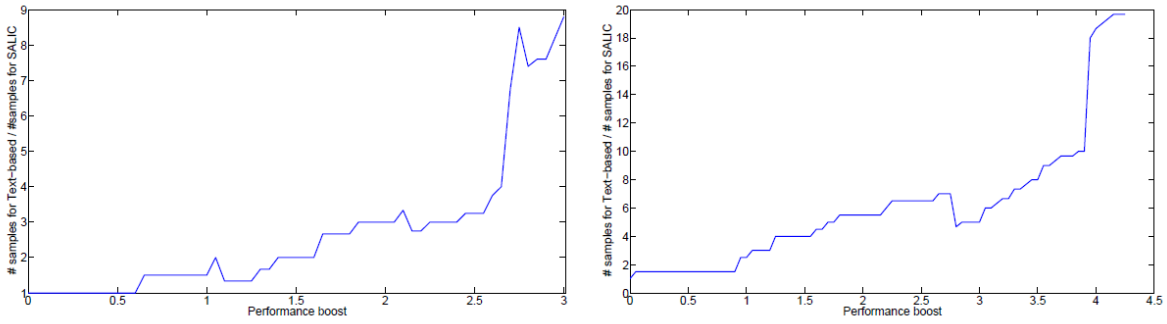


(a) Fisher-based features. (b) CNN-based features.

Fig. 4-3. Comparing with fusion baselines.

Why Active Learning when Tags are Free

In this section we want to demonstrate the gain in scalability achieved by actively forming the training set compared to adding positive and negative images in an informativeness-agnostic manner. For this experiment, the text-based approach plays the role of the informativeness-agnostic learning algorithm. For demonstrating the gain in scalability, we compute the ratio of the images that are required by the text-based approach to the images that are required by SALIC in order to achieve the exact same performance. Fig. 4-4 displays the plot of this ratio with respect to the achieved performance boost. For the CNN-based features, in order to achieve a 4% boost in mAP, the text-based method requires 20 times more images than SALIC (Fig. 4-4b). Similarly, for the Fisher-based features, the informativeness-agnostic approach requires 9 times more instances to reach the same performance as SALIC for a 3% boost on its performance (Fig. 4-4a). The fact that we need an order of scale more data instances to achieve the same performance gain, shows the importance of active learning even in the case where labels can be obtained for free. Moreover, it is interesting to note that the text-based approach was not able to reach SALIC's maximum performance, even after 200 iterations (within the first 50 iterations, it approximately achieved half of the performance boost compared to SALIC as it can be seen in Fig. 4-3).



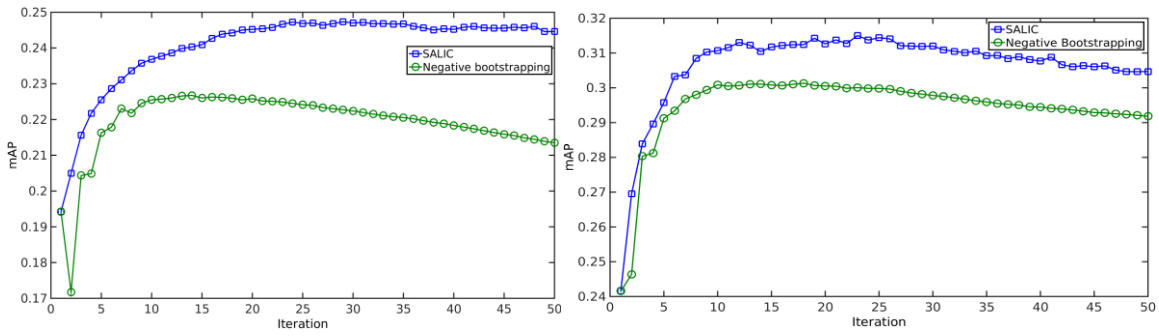
(a) Fisher based features. (b) CNN based features.

Fig. 4-4. Ratio of samples, required to achieve the same performance between the text-based approach (i.e. random) and SALIC.

Comparing with State-of-the-art Methods

Comparing with Active Learning: Here we compare SALIC with the negative bootstrapping approach presented in [Li et al., 2013] (i.e. selecting in each iteration the 100 negative images that are most misclassified and utilizing the model aggregation approach to cope with the class imbalance problem). The results can be seen in Fig. 4-5, for the Fisher- and the CNN-based features,

respectively. We see that SALIC continuously increases the performance of the classifiers and constantly outperforms the other approach in both cases.



(a) Fisher-based features. (b) CNN-based features.

Fig. 4-5. Comparing with Negative Bootstrapping [Li et al., 2013].

Comparing with Weakly Supervised Learning (WSL): This section discusses a comparison with a weakly supervised method that selects images by querying search engines [Papadopoulou & Mezaris, 2015]. The authors base their approach on the observation that search engines tend to provide accurate results for the top images. The proposed approach constructs a set of queries by i) translating the original query to 15 different languages, ii) using hyponyms, hypernyms and synonyms from WordNet, and, iii) finding related terms within the results of the Google text search engine. Then, they query image search engines (i.e. Google, Bing and Flickr) with each term in the previously constructed set and retrieve the top 24 images of each query[9]. In order to compare SALIC with this approach, we applied our method to the dataset used in [Papadopoulou & Mezaris, 2015], which consists of 40 ImageNet concepts. As there was no manually labelled set for these concepts, we selected the initial 100+100 images to train H_0 from the MIRFLICKR-1M dataset using the text-based approach. It is evident from the results (see Fig. 4-6) that SALIC significantly outperforms the weakly supervised approach for 30 out of the 40 concepts. Moreover, for 3 of the concepts, namely "*animal*", "*rhino*" and "*vehicle*", SALIC fails to gather good quality training samples. For the first two concepts this can be attributed to their generic nature, since they include many diversiform sub-concepts (e.g. dogs, cats, insects, fish, etc. for animal and boats, buses, cars, airplanes for vehicle). For the case of "*rhino*", there are not enough images depicting this concept in MIRFLICKR-1M, and it would require a larger pool of candidates in order to gather a better training set.

[9] 24 was found to be optimal in [Papadopoulou & Mezaris, 2015].

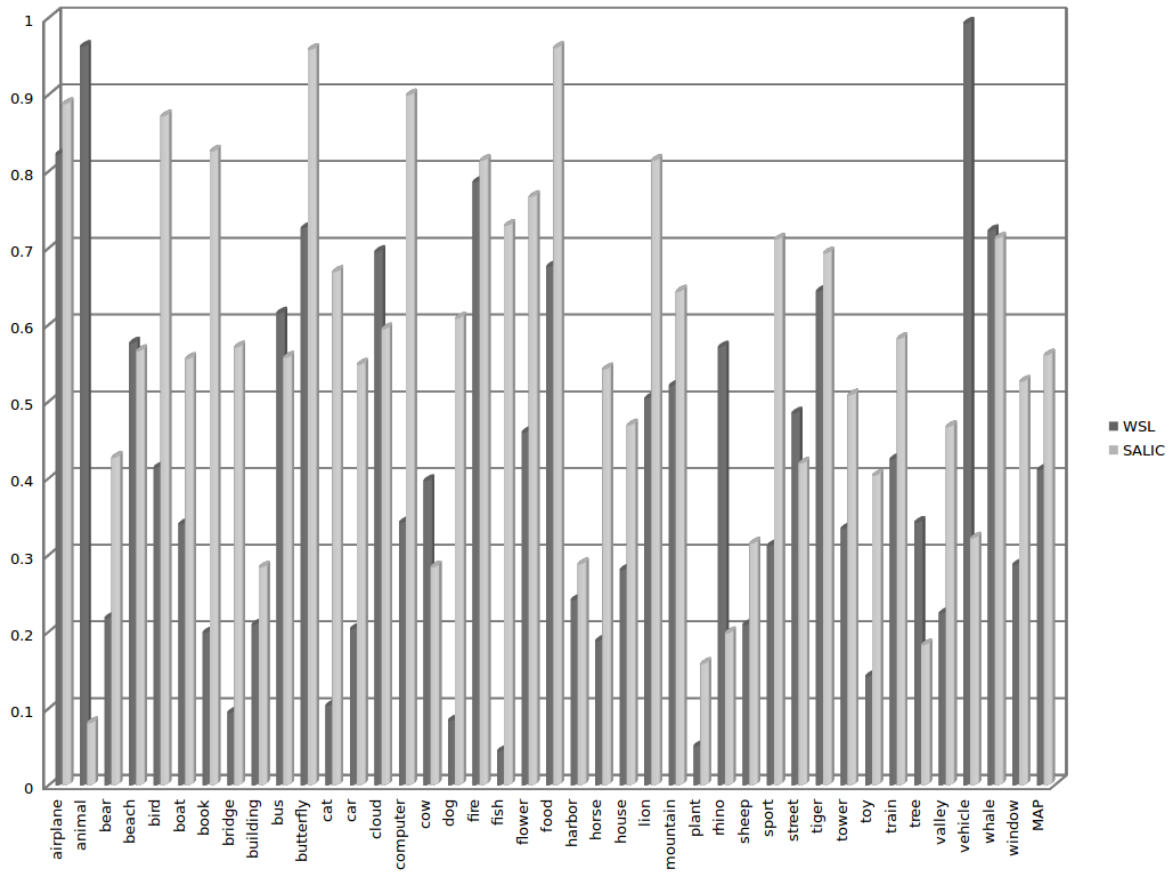


Fig. 4-6. Comparing with weakly supervised learning.

In this section, we presented SALIC, an automatic active learning approach for image classification, where the oracle is replaced with web users and the pool of candidates with user tagged images. The contribution of SALIC is a probabilistic approach for joint maximization of the samples' informativeness and the oracle's confidence. Experimental results show the superiority of SALIC compared to baselines and state-of-the-art approaches. Moreover, we argue that despite the abundant availability of user tagged images that can be labelled with no cost, the existence of methods like SALIC are necessary to ensure the scalability of image classification applications to the constantly increasing demands. The full set of experiments for this work can be found in [Chatzilari et al., 2015].

4.1.2. Scalable Image Classification Using Product Compressed Sensing and Sparse Represantations

Recently, considerable effort has been invested to bridge the **semantic gap**, i.e., the gap between the semantic descriptions that humans and computers assign to objects. Towards this direction, sparse representations are shown to convey important semantic information that can be extracted and used in diverse challenging computer vision and pattern recognition application domains. Furthermore, combining sparse representations with appropriate discriminant criteria may lead to performances superior to that of the state-of-the-art discriminative algorithms in classification problems. In the following we review a few indicative examples in this direction.

Semantic issues have been tackled using sparse image reconstruction based approaches in object recognition [Wang et al., 2009]. An image is encoded using patch-specific Mixture Gaussian Models

and reconstructed as linear combinations of a number of training images containing the objects in question. The reconstruction coefficients are inherited by the corresponding labels providing annotation to the input image. Towards this direction, the problem of accelerating sparse coding based scalable image annotation has also been addressed [Huang et al., 2013]. The pursuit of an accurate solution is based on an iterative bi-level method, which actually reduces the large-scale sparse coding problem to a number of smaller sub-problems.

Unlike most feature extraction algorithms that have been proposed for dimensionality reduction, Yang et al. (2010) propose a CS-based method, which combines sparse representation with feature selection. An objective function involving a matrix that weights the features according to their discriminant power is minimized in a sparse reconstruction fashion by adopting an OMP-based algorithm. In the same work, a two-stage hierarchical architecture using directional and adaptive filters for feature detection has been implemented. The proposed methodology has been applied and evaluated in the pedestrian detection problem proving its efficiency in real-world problems.

Sparsity, Reconstruction error and Discrimination power are combined in a common objective function in [Huang & Aviyente, 2006]. A hybrid approach that combines the discrimination power of discriminative methods with the reconstruction capacity of sparse representation has been proposed in an attempt to handle corrupted signals. However, there is a trade-off between the two aforementioned aspects, which stems from the potential noise, the completeness of the data, the number of outliers, etc. that must be effectively handled. Even in the case where the data are noise-free and the discriminative methods are more robust, it has been experimentally shown that, by combining the two methods, superior performance is obtained.

Compressed Sensing has also been utilized for recovering the sparse foreground of a scene as well as the silhouettes of the foreground objects. In [Cevher et al., 2008], based on the CS representation of the background image, the authors propose a method devoid of any foreground reconstruction for performing background subtraction and object detection, using convex optimization and greedy methods. The authors also recover the silhouettes of foreground objects by learning a low-dimensional CS representation of the background image, robust against background variations.

Sparse representations with over-complete dictionaries have also been applied on image denoising [Elad & Aharon, 2006]. Utilizing the K-SVD technique, the authors train dictionaries based on a set of high-quality image patches or based on the patches of the noisy image itself. The image is iteratively de-noised through a sparse coding and a dictionary update stage. Based on a similar approach, K-SVD has been utilized for the restoration of images and video [Mairal et al., 2007]. In this work, the sparse representations are obtained via a multi-scale dictionary learned using an example-based approach. Finally, the above de-noising algorithm has also been extended for color image restoration, demosaicing and inpainting [Mairal et al., 2008].

Although CS is recognized in literature for its ability to compactly represent digital data [Candès & Walkin, 2008], in this deliverable we investigate the extent to which CS-based methods can be used for discrimination purposes in a scalable manner. More specifically, our aim is to identify potential ways for exploiting the CS theory in large-scale image annotation. The rationale of our work is based on the outcome of some important works like [Wright et al., 2009], which claim that sparse representations have an inherent discriminant nature. Indeed, using a small subset among a pool of features offers a compact way to represent an image and therefore the selected features strongly characterise this image. Based on the above claim, we envisage that **combining sparse representation with discriminative methods** may lead to performance competitive to that of the state-of-the-art in large-scale classification problems [Huang & Aviyente, 2006; Yang et al., 2010].

In this deliverable, we rely on the principles of CS to propose a novel scalable **Product Compressive Sampling (PCS)** method for dimensionality reduction in the image annotation domain. PCS decomposes a high-dimensional vector into a number of smaller vector segments performing an equal number of CS projections and concatenating the results. Through both a theoretical analysis

and an experimental comparison with typical CS, we show that PCS exponentially reduces the computational load of the dimensionality reduction process, while at the same moment displays performance equivalent to CS. We further establish a connection between the performance of PCS and the level of the data sparsity. Finally, a comparison with the state-of-the-art, shows that our method displays competitive performance in terms of classification performance, while at the same time outperforms the state-of-the-art methods in terms of computational efficiency.

4.1.2.1. Product Compressed Sensing

As described above, PCS decomposes a high-dimensional vector into a number of sub-vectors and applies CS reducing the dimensionality of each sub-vector. The resulting reduced sub-vectors are then concatenated providing the final low-dimensional vector. More specifically, let $\mathbf{x} = [x_1, \ldots, x_n]^T$ be an *n*-dimensional vector and *m* the desired reduced dimensionality. At the first stage, the components $x_j, j \in \{1, 2, \ldots, n\}$ are sorted in ascending mode. To keep the notation as simple as possible, let us consider the above **x** as the sorted vector. Subsequently, **x** is subdivided into a set $\{\mathbf{x}_i\}_{i=1}^{b}$ of *b* vector-blocks of size $q = \frac{n}{b}$ each, with $\mathbf{x}_i = [x_i, x_{b+i}, x_{2b+i}, \ldots, x_{(q-1)b+i}]^T$ for $i = 1, 2, \ldots, b$. Having obtained the blocks, PCS reduces the dimensionality *q* of each block to $\frac{m}{b}$ and subsequently concatenates the resulting vectors to form a $b \times \frac{m}{b} = m$-dimensional output vector.

In the above approach, it is clear that $\frac{n}{b}$ and $\frac{m}{b}$ must be integers, since the former expresses the length of each block constructed from the initial vector, while the latter is the length of each block after the reduction. Provided that the above limitation is satisfied, a random projection matrix $\mathbf{D}$ of size $\frac{m}{b} \times \frac{n}{b}$ is firstly orthonormalized through an SVD step and then used for reducing the data of each block $\mathbf{x}_i$ to $\frac{m}{b}$ dimensions, producing $\mathbf{y}_i$, for $i = 1, 2, \ldots, b$, through $\mathbf{y}_i = \mathbf{D}\mathbf{x}_i$.

The finally reduced vector is built by concatenating the obtained blocks: $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_b]$ which consists of *m* dimensions. A block diagram of the proposed PCS approach is given in Fig. 4-7. At this point, it is worth noting that in the extreme case where $b = 1$, PCS collapses to the typical CS method, since in this case the whole vector is considered as one block.

The functionality of PCS relies upon the data sparsity assumption. In fact, when subdividing a sparse vector we expect that, on average, the vector segments are also supposed to commensurately inherit the sparsity of the main vector, therefore permitting the use of random projections per each segment. As a matter of fact, let us consider a feature vector containing $k$non-zero values. Then, as already mentioned, the lower bound m_L of the reduced dimensionality for the initial high-dimensional vector must be on the order of $O(k \cdot log(\frac{n}{k}))$. Assume now that the initial vector is decomposed into $b$equal length blocks employing the proposed approach. From the way the several blocks are constructed, the sparsity of each block is expected to be on average $\frac{k}{b}$, since actually the above block decomposition procedure comprises a uniform subsampling of the initial sorted vector. Consequently, the corresponding lower bound m'_L for each block becomes:

$$m'_L = O(\frac{k}{b} log(\frac{\frac{n}{b}}{\frac{k}{b}})) = \frac{1}{b} O(k \cdot log(\frac{n}{k})) = \frac{1}{b} m_L$$

which after the concatenation phase is aggregated to $b \cdot \frac{1}{b} m_L = m_L$. In short, this means that PCS maintains on average the dimensionality bounds of CS in the reduction process.

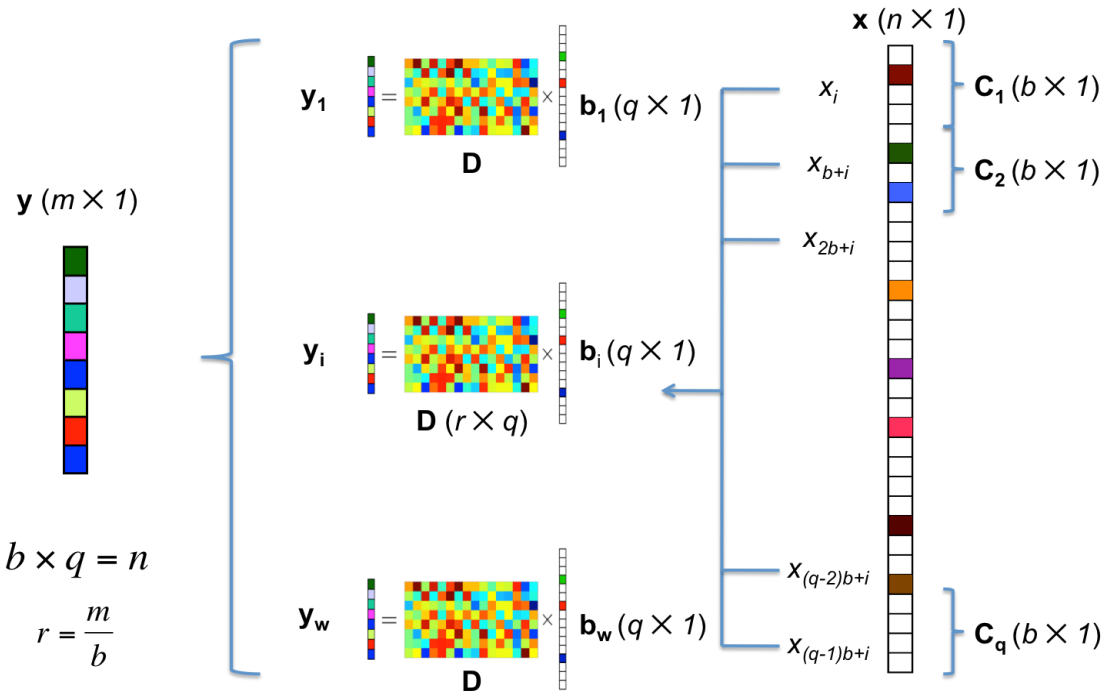


Fig. 4-7. Block diagram of PCS.

Computational Analysis: The computational load using CS consists of performing SVD on a random matrix along with a number of matrix multiplications during the projection phase. The number of calculations involved in the above process is exponentially associated with the number of dimensions. Based on this, PCS earns its advantage by splitting the initial CS problem to a number of smaller CS sub-problems by decomposing a vector into an equal number of blocks. In this way, PCS achieves to alleviate the computational load of CS in the way explicitly described in the following analysis.

The computational complexity of performing SVD on an $m \times n$ matrix is on the order of $4m^2n + 8mn^2 + 9n^3$. Similarly, the computational complexity of multiplying an $m \times n$ matrix by an $n \times 1$ vector is on the order of $\mathcal{O}(mn)$. In order to compare PCS with CS, for the sake of simplicity, let us investigate the case where the number of dimensions in the reduced space is $\frac{n}{2}$. Under these circumstances, using CS requires SVD on a $n \times \frac{n}{2}$ random matrix with computational complexity $\mathcal{O}(5n^3)$, and a multiplication between the transpose of the resulting matrix by the initial $n \times 1$ vector, which is $\mathcal{O}(\frac{n^2}{2})$. On the other hand, using PCS with b blocks, requires SVD calculation of a matrix of size $\frac{n}{b} \times \frac{n}{2b}$, which is $\mathcal{O}(\frac{1}{b^3}5n^3)$, that is $\mathcal{O}(b^3)$times less than using CS. Subsequently, the resulting projection matrix is multiplied by the b vector-blocks, which is $\mathcal{O}(\frac{1}{b^2}\frac{n^2}{2})$, that is $\mathcal{O}(b^2)$times less than using CS.

The previous analysis, clearly shows that the computational benefit using PCS is exponentially associated with the number of blocks used. However, this might come at the cost of a loss of information, since by significantly shrinking the length of the blocks might cause the loss of vector-structure. This uncertainty led us to investigate the performance of PCS as a function of the number of blocks used, through an experiment presented in Section 4.1.2.2. As we will see, the results show that the performance of PCS proves to be independent of the number of blocks used.

4.1.2.2. Experimental Results

We conducted a series of experiments in order to investigate the potential of PCS as a dimensionality reduction technique in the problem of image annotation in large-scale datasets. For training the SVM classification models, the implementation of LibSVM was used [Chang & Lin, 2011]. The mean Average Precision (mAP) served as the evaluation metric. All experiments were run on a 12 core Intel Xeon (R) CPU ES-2620 v2 @ 2.10 GHz with 128 GB memory.

For the experiments we have used the benchmarking dataset of the PASCAL VOC 2012 competition [Everingham et al., 2012]. The dataset consists of 5717 training and 5823 test images collected from flickr. The images are annotated with 20 concepts in a multi-label manner. Applying the feature extraction procedure (cf. Chapter 3) on the datasets, we came up with a set of 11540 feature vectors each of 327680 dimensions, which require approximately 14 GB of memory using single floating format. Such excessive memory requirements are clearly intractable for many practical reasons.

Investigating PCS robustness as a function of the number of blocks:

Although in Subsection 4.1.2.1 it has been theoretically shown that increasing b benefits the computational complexity of the PCS process, the question how b can affect the classification performance using PCS is still open. In an attempt to answer this question, we set the number of reduced dimensions to 163840, i.e. half of the initial dimensionality of the data, and we varied the number of blocks b in the range of 1, 2, 2^2,… , 2^9. For each setting, we counted the time elapsed during the dimensionality reduction process and we calculated the classification performance using PCS. The results are collectively illustrated in Fig. 4-8. The horizontal axis depicts the number of blocks used. The left vertical axis depicts the mAP, while the right vertical axis depicts the computational time required for reducing the dimensionality of the training data using PCS. The latter includes the time required for both the SVD calculation and the matrix multiplication processes.

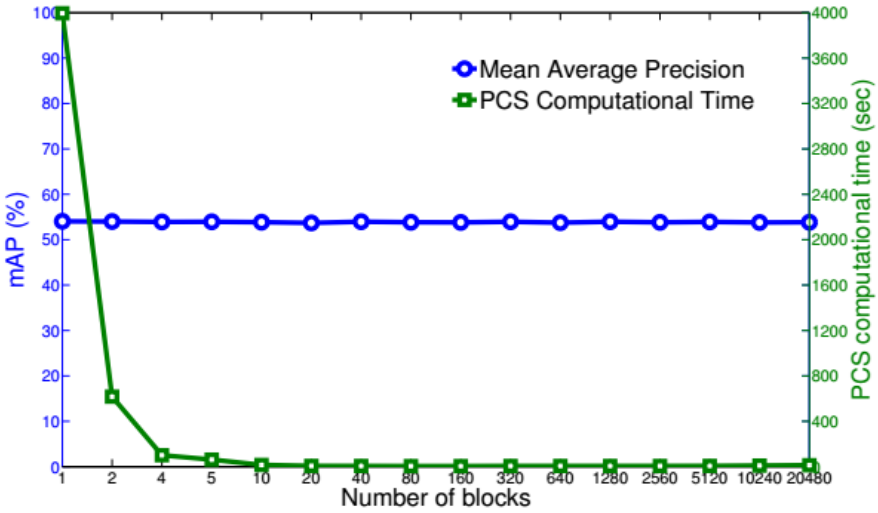


Fig. 4-8. Mean average precision and training computational time versus number of blocks used in PCS.

Observing the "PCS computational time" curve in Fig. 4-8, it is clear that, as the number of blocks increases, the computational complexity decreases exponentially. This result can be associated with the theoretical findings, verifying the computational analysis presented in Subsection 4.1.2.1. Furthermore, interestingly, the robustness of PCS using different numbers of blocks is evident (see Fig. 4-8, "Mean average precision" curve). This finding allows for using the maximum possible number of blocks, since it offers the minimum computational load, while it does not deteriorate the

classification performance. Closing this subsection, it is worth recalling that PCS for *b* = 1, i.e. using one block, collapses to CS (Subsection 4.1.2.1).

Comparing the classification performance of PCS versus CS:

Our next concern was to investigate the classification performance of PCS, as a function of the number m of the reduced dimensions, compared to that of CS. For this purpose, a series of consecutive experiments was carried out, where at each iteration we set $m = 10 \cdot 2^p$ and we varied the exponent p in the range from 1 ($m = 20$) to 11 ($m = 20480$), while maintaining the remaining settings unchanged. In this context, using different integers for m, we employed both CS and PCS and we projected the initial high-dimensional data into the corresponding m-dimensional space. The above projected data were subsequently fed into SVM for classification. The classification performance results are illustrated in Fig. 4-9 for both CS (green rectangles) and PCS (blue circles). The reduced number of dimensions is depicted in the horizontal axis, while the mAP is depicted in the vertical axis. For comparison reasons, the baseline mAPusing SVM on the initial data, with no dimensionality reduction is also depicted with the solid red line.

A couple of important remarks could be drawn from Fig. 4-9. First, it is strongly evident that the two curves corresponding to CS and PCS are almost identical, proving the equal classification performance of PCS compared to CS. This finding combined with the computational complexity advantage of PCS (Subsection 4.1.2.1) proves its superiority over CS. In the same vein, notice that although the initial dimensionality of the data is 327680, the maximum reduced dimensionality was set to 20480. The reason is that for larger dimensions CS was infeasible due to the orthonormalization of a very big matrix, which reinforces the advantage of PCS versus CS. Second, it is clear that the mAP for both CS and PCS is ever increasing as a function of the number of the reduced dimensions. Moreover, a closer inspection of Fig. 4-9 interestingly shows that the mAP starts to converge satisfactorily to the baseline in the vicinity of the theoretically required number of dimensions, i.e. 6464, which can be associated with the lower boundary analysis presented in subsection 4.1.2.1.

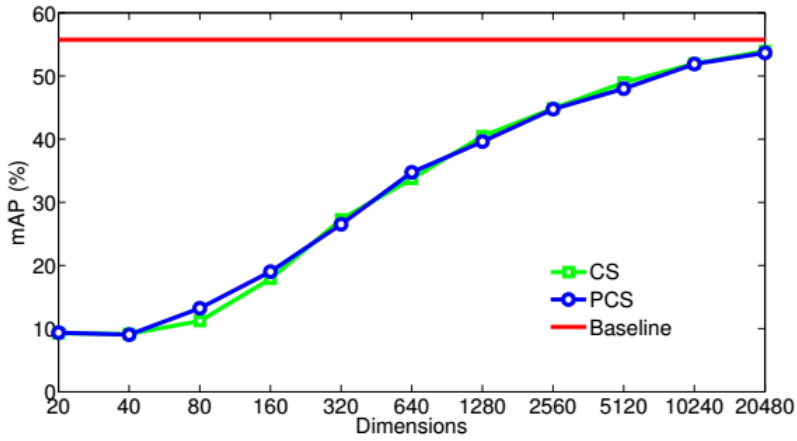


Fig. 4-9. Mean average precision using PCS in a range of different dimensions.

PCS vs Random Feature Selection and Dependence of PCS on Data Sparsity:

A reasonable question that might arise so far is why not use a random feature selection (RFS) approach instead of calculating random linear combinations of the initial features and how does the data sparsity – and consequently compressibility – affect the performance of random projections. In an attempt to answer these questions, we propose a methodology for comparing the performance of PCS with RFS as a function of the sparsity level of the data. The methodology is based on "sparsifying" the original dataset using six different threshold values and investigating the robustness of the two

above methods in the resulting artificial datasets. From a practical point of view, sparsifying the data is supposed to deteriorate the classification performance of both methods, since it leads to considerable loss of information. However, from this artificial experiment, we expect that a number of important findings regarding the effect of the data sparsity on random projections can be derived.

For each sparsity level and dimensionality we calculated the difference between the mAP's obtained by using PCS and RFS and we estimated the percentage gain in classification performance obtained by PCS over RFS. The results are illustrated in Fig. 4-10. The x-axis depicts the number of reduced dimensions. The y-axis depicts the percentage of mAP gain using PCS versus RFS. Each curve corresponds to a different sparsity level, as this is expressed by the percentage of zeros contained in the sparsified data. From Fig. 4-10, it is evident that the gain in performance using PCS instead of RFS increases as the sparsity level increases too and as the number of dimensions decreases. This improvement of performance reaches 80% in 1280 dimensions with 97.72% sparsity, highlighting the robustness of PCS versus RFS in low dimensions on sparse data.

Intuitively, at a first glance, there is seemingly nothing special about random projections (e.g. PCS) against RFS, due to the random nature of both. However, random projections clearly take into account all the initial data features, while in contrast, selecting a number of specific features inevitably avoids the rest leading to considerable loss of information. This advantage provides credibility to PCS as a smart dimensionality reduction method over other naive random feature selection schemes under the data sparsity assumption.

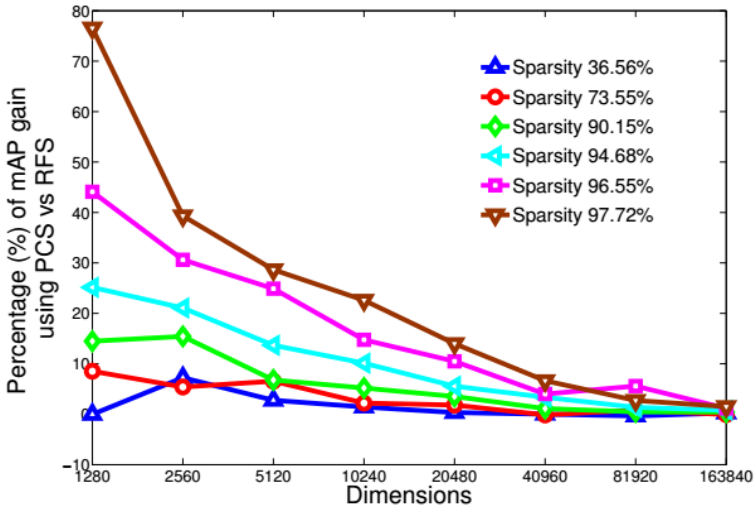


Fig. 4-10. Percentage (%) of mAP gain using PCS vs RFS.

Comparison with the state-of-the-art:

In this section, we compare the performance of PCS with the state-of-the-art PQ [Jégou et al., 2011] and PCA [Jolliffe, 2002]. Regarding PQ, the experiments were carried out using the initial 327680 dimensional data. However, since the application of PCA on the 327680 dimensional data was infeasible (see PCS vs PCA comparison in the following), the experiments involving PCA were conducted using only one out of the eight spatial pyramids of the data. That is, the dimensionality of the data for this particular set of experiments was 40960. For this purpose, the results are separately presented for PQ and PCA.

PCS vs PQ: For PQ we have used the k-means implementation presented in [Jégou et al., 2011]. The theoretical complexity of this implementation is $O(lkNn)$, where k is the number of the prototype centroids trained by K-means, l is the number of iterations, n is the initial dimensionality (327680 in

our case) and *N* is the number of training samples (5717 in our case). The dependence of the above complexity on *I*, *k* and *N* does not allow a direct comparison with the corresponding of PCS. However, since in the literature it has been shown that a considerably large number of all these three parameters is needed [Jégou et al., 2011], it is clear that PQ might face severe computational difficulties during its application. In our experiment, we set *I* = 100 and *k* = 256, as proposed in [Jégou et al., 2011]. Using these settings, in the following we provide a comparison between PCS and PQ from both a classification performance and a computational complexity point of view, where the latter has been based on the CPU times during the experiment.

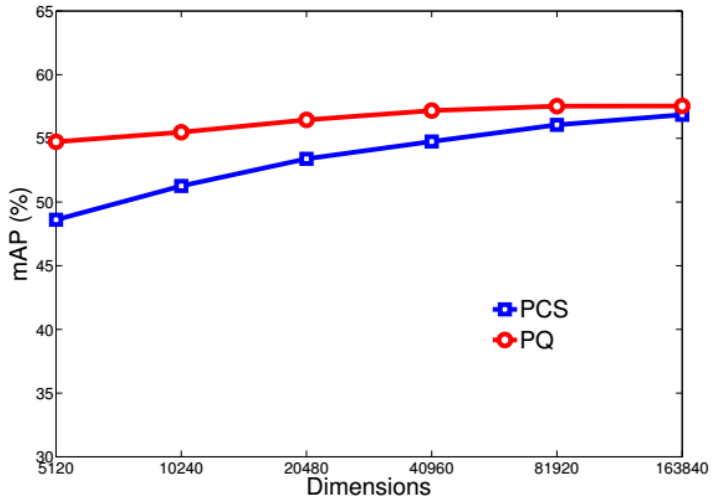


Fig. 4-11. Comparison between PCS and PQ.

The classification performance comparison between PCS and PQ is illustrated in Fig. 4-11. The mAP is depicted in the y-axis, while the number m of reduced dimensions is depicted in the x-axis. Based on the result of Section IV-A on the least required number of reduced dimensions, we allowed *m* to take values in the range from 5120 to 163840 dimensions. From Fig. 4-11, it is clear that although PQ outperforms PCS, the difference between the two becomes negligible at high dimensions. Fig. 4-12 jointly illustrates the percentage loss in performance (left vertical axis) and the corresponding speedup (right vertical axis) using PCS versus PQ. Interestingly, from Fig. 4-12, the loss in performance at 163840 dimensions is only 1%, which is a difference of 0.69 units of mAP, while PCS provides a 362 times speed-up over PQ. The above loss in performance becomes approximately 11% at 5120 dimensions, while PCS is 23 times faster than PQ at the same dimensions. Such differences in computational efficiency, in conjunction with the corresponding small compromise in terms of classification performance, provide a huge potential of PCS as an effective dimensionality reduction method in large-scale classification problems.

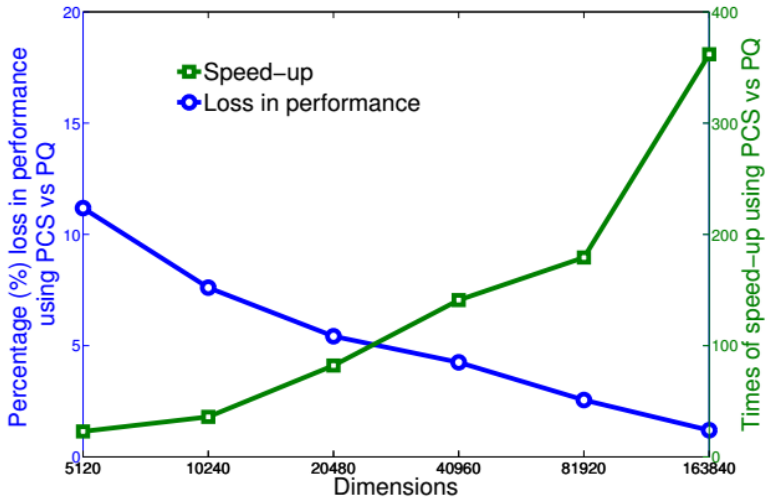


Fig. 4-12. Comparison between PCS and PQ.

PCS vs PCA: The mAP results from the comparison between PCS and PCA are plotted in Fig. 4-13. Since in this case the initial dimensionality of the data was 40960, the number of reduced dimensions was allowed to take smaller values down to 1280. Similarly to the comparison with PQ, from Fig. 4-13, we observe that although in low dimensions, the superiority of PCA over PCS is evident, by increasing the number of dimensions, PCS exhibits classification performance competitive to PCA. The robustness of PCA in low dimensions, is well justified by the fact that PCA by definition attempts to encode the data information into the least possible eigenvectors. However, this advantage comes at the cost of excessive computational and memory requirements. More specifically, regardless of the number of reduced dimensions, PCA requires the computation and eigenanalysis of the data covariance matrix, which is on the order of $n^2N + N^3$, where n is the dimensionality and N is the number of the samples. The above computational complexity in conjunction with the excessive memory requirements may render the application of PCA prohibitive. As a matter of fact, in this particular experiment, it was infeasible to apply PCA on the data consisting of all eight spatial pyramids (327680 dimensions), since 400 GB of memory was required only for storing the covariance matrix. This is the reason why, as stated previously, we used only one out of the eight spatial pyramids.

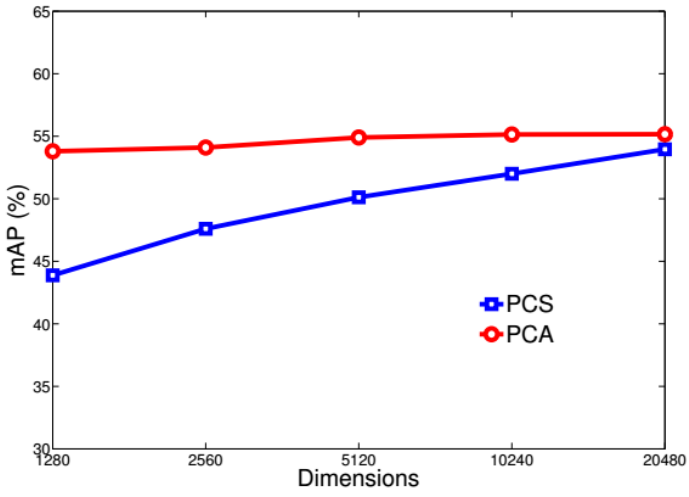


Fig. 4-13. Comparison between PCS and PCA.

On the other hand, CS requires the construction of an $m \times n$ random matrix, where m is much smaller than n, and the SVD of this matrix during the basis orthonormalization step. Furthermore, using PCS, the SVD is decomposed into a number of smaller SVD steps, leading to a computational load orders of magnitude less than the corresponding of PCA. As a matter of fact, a comparison analogous to Fig. 4-12 is illustrated in Fig. 4-14. Although the loss in performance using PCS instead of PCA reaches 18% at 1280 dimensions, PCS is more than 300000 times faster. Moreover, the above loss in performance reduces to only around 2% at 20480 dimensions, while the corresponding speed-up using PCS is around 24000.

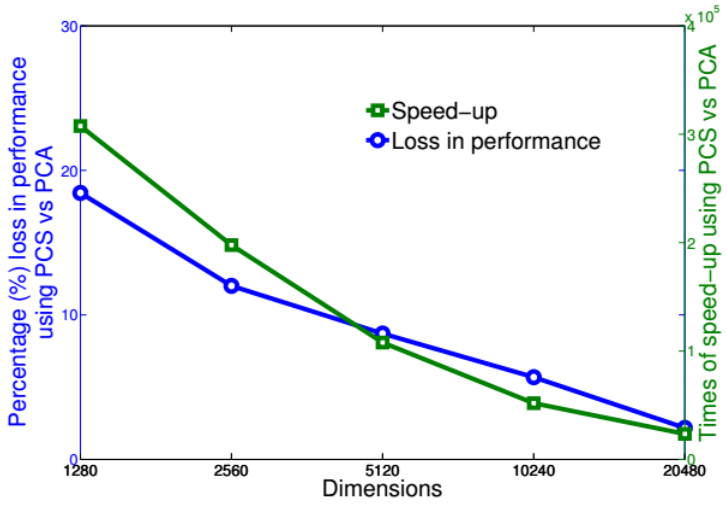


Fig. 4-14. Comparison between PCS and PCA.

Summarising the results of the comparison between PCS, PQ and PCA, the deterioration in performance of PCS when reducing the number of dimensions can be attributed to the democratic nature of the reduced dimensions, which postulates a representative number of dimensions in order to maintain the important information. It must be emphasised though that the profit in terms of computational complexity using PCS compensates for the corresponding loss in performance, which for many practical reasons can be proven negligible.

4.1.3. Acceleration of Context-dependent Quantum Methods for Semantic Media Classification

We pursue several directions to identify quantum-like patterns in semantic media classification. The common element to these is that theoretically, many such patterns manifest at the level of observed correlations. These correlations can be the familiar classical correlations, but also non-local correlations enabled by the mathematical formalism of quantum mechanics – these emerge from contextual behaviour. Furthermore, correlations may not be restricted to the quantum set, and we are keen to see whether we can identify correlations stronger than what is allowed by the quantum mechanical description. This typology of correlations underlying advanced access could result in a new preservable as part of the appraisal scenario, changing the very foundation that the computational outcomes are resting on.

As mentioned in the previous chapter, for a practical analysis of context-dependent correlations, we have been developing a high-performance qualitative machine learning algorithm called Somoclu. This tool is primarily meant for training extremely large emergent self-organizing maps on supercomputers, but it is also the fastest implementation running on a single node for exploratory data analysis. Whereas below we refer to conceptual and scalability results demonstrated on texts as

one form of semantic medium, the approach is generic and applies to any features represented by vectors, including image descriptors. Its multipurpose nature means that it can be used to analyze concept and semantic drifts (the focus of T4.4) and to uncover hidden correlations in sparse data collections.

Beyond that, to augment Somoclu and for analysing the typology of observed correlations, we developed a tool called **Ncpol2sdpa**[10]. Ncpol2sdpa is less applied and it is more of a research tool – its purpose is to verify a hypothesis of quantum-likeness. It can bound certain convex sets, which we use to justify our hypotheses of quantum-like behaviour in digital collections. For instance, we can exclude hidden variable theorems in networked data, or verify the strength of observed correlations. It is written in Python and while it is not a high-performance computing tool in itself, it acts as a necessary preparatory step to solve extreme-scale semidefinite programs on accelerators – the results of these computations yield the bounds we seek.

Both tools will be deployed shortly on the Tate metadata[11] as a form of other text-based environmental information associated with the objects, such as documentation created during their creation and curation. We also had plans to analyze clickstreams from museums exemplifying changing patterns of information seeking behaviour in interaction with semantic content but, as such data are hard to come by yet, they will be replaced by similar non-museum data less prone to privacy issues for simulated but similar results.

To connect the three key concepts in the subtitle, i.e. acceleration and scalability, context-dependent quantum-inspired computational methods, and the problem of DO classification based on semantic content, we start with a brief overview of quantum-like systems. In the next step, we briefly outline how a key concept in classification by machine learning, that of energy minimization, leads to a new approach which considers semantics as "energy" in a metaphoric sense[12]. From this it follows that just like energy is stored in fields in physics to induce change, we too can conceive semantics in a field form. The proof for this line of thought is the scalability and generic applicability of Somoclu as a field analytical tool to the classification of semantic media. However, currently it must be left open if classical physics is sufficient to model evolving semantic systems, or one can detect non-classical system behaviour in the digital world as well.

The Concept of Quantum-likeness

Quantum-like systems are those to which quantum mechanical (QM) formulae are applicable and bring interpretable results. The idea came as a surprise in the 90s and was shown to hold in domains like economy, evolutionary biology and cognition. These fall outside of the subatomic world where the rules of QM hold, whereas in the atomic world and beyond, including human societies, classical (Newtonian) mechanics (CM) was supposed to exclusively rule. The applicability proposal goes back to [Aerts & Gabora, 2005; Khrennikov, 2010], with many results showing that genuine concepts of QM such as *contextuality* (the dependency of results on the observation/readout sequence) and *entanglement* (being at two locations at the same time) can be used to model human behaviour and its products. This applicability goes back to an alternative interpretation of probability [Khrennikov, 2010] and a typology of correlations according to their strength, coming from mathematics [Tsirelson, 1980; Gisin, 2009]. Obviously, if correlations in the data can be of several kinds, then the

[10] Ncpol2sdpa is freely available under GNU GPL at https://goo.gl/3c0gNO
[11] We expect to get results from the ongoing experiments by March.
[12] Mathematical "energy" and machine learning (ML) are related, the latter often being based on minimizing a constrained multivariate function such as a loss function. Concepts in feature space "sit" at global energy minima, representing the cost of a classification decision as an energy minimizing process. This suggests that ML must identify concepts with such minima, and since energy in physics is carried by a field or a respective topological mapping, concepts naturally have something to do with energy as work capacity.

results derived from them must be different too, an implication of vast importance for machine learning, and any field using it for the automatic categorization of digital objects, including DP.

With the increasing number of evidence for the above, latest research is taking two directions: investing into **quantum machine learning (QML)** [Wittek, 2014], and exploring the implications such as **quantum information science** [Bawden et al., 2015]. For digital libraries, the question arises if semantic content in media partly "behaves" according to CM, or is partly also quantum-like. To explore the viability of this observation, we focused on two sets of activities for this deliverable, developing a field theory of semantic content as a common denominator, and designing tools to detect field-like vs. quantum-like content behaviour.

Accelerating Scalable Machine Learning for Supercomputers

Next we discuss the scalability and acceleration aspects of Somoclu as part of tool development for the observation of evolving semantic content in vector space.

Background

Although supercomputing is not a mainstream approach to handling DOs for DP yet, over time one can expect significant changes in this respect. For example graphics processing units (GPUs) were originally designed to accelerate computer graphics through massive on-chip parallelism. As the inherent data-parallelism in graphics applications is also apparent in many other fields, GPUs have evolved into powerful tools for more general cases of computationally intensive tasks. For instance, researchers have studied how GPUs can be applied to problem domains such as scientific computing [Owens et al., 2007; Ufimtsev & Martinez, 2008; Stone et al., 2010], and visual applications [Strong & Gong, 2008; van de Sande et al., 2011; Daróczy et al., 2011]. These applications often rely on distributed nodes to overcome the limitations of device memory. This paradigm is known as **massively** or **embarrassingly parallel computation**. Only a handful of data mining applications benefit from such infrastructure, yet, it is a natural way to address both images and texts by the same computational technology.

Since the initial steps of text mining are typically data-intensive, and the ease of deployment of algorithms is an important factor in developing advanced applications, we introduced a flexible, distributed text mining workflow that performs I/O-bound operations on CPUs with industry-standard tools and then runs compute-bound operations on GPUs which are optimized to ensure coalesced memory access and effective use of shared memory. Such heterogeneous computing aims to combine the parallelism of traditional multi-core CPUs and GPU accelerator cores to deliver unprecedented levels of performance [Brodtkorb et al., 2010]. While the phrase typically refers to a single node, a distributed environment may be constructed from such heterogeneous nodes.

CPUs excel in running single-threaded processes, or in multithreaded applications in which a thread often consists of fairly complicated sequential code. Graphics processors are ideally suited for computations that can be run in parallel on numerous data elements simultaneously. This typically involves arithmetic on large data sets, where the same operation can be performed across thousands of elements at the same time. This is actually a requirement for good performance: the software must use a large number of threads. The overhead of creating new threads is minimal compared to CPUs that typically take thousands of clock cycles to generate and schedule, and a low number of threads will not perform well on GPU [Kirk & Hwu, 2009]. The decomposition and scheduling of computation among CPU cores and GPUs are not trivial even on a single node [Jiménez et al., 2009; Lee et al., 2009; Luk et al., 2009], and the task is even more complicated for computer clusters [Phillips et al., 2008]. In order to issue work to several GPUs concurrently, a program needs the same number of CPU threads, each with its own context. All inter-GPU communication takes place via host nodes. Threads can be lightweight (pthreads, OpenMP, etc. [Kuhn et al., 2000]) or heavyweight (MPI [Koop et al., 2006]). Any CPU multi-threading or message-passing API or library can be used, as CPU

thread management is completely orthogonal to GPGPU programming. For example, one can add GPU processing to an existing MPI application by porting the compute-intensive portions of the code without changing the communication structure [Nvidia, 2014]. However, the efficient utilisation of all CPU and GPU cores remains an open question.

High-performance computing (HPC) uses supercomputers and computer clusters to solve advanced computational problems. A supercomputer is purpose-built hardware which is typically very costly to build. Clusters combine powerful workstations or even commodity hardware through a high-speed network to achieve higher scales. Since even commodity hardware is extremely powerful these days, enormous clusters have been overtaking supercomputers in the rankings of computational performance for the past decade.

The computational expense to execute data or text mining based analysis as advanced services in real-world applications such as digital libraries has been identified as the major cause for the lack of more widespread use of such services [Sanderson & Watry, 2007].

The initial step of text mining translates the unstructured documents to an algorithmically more manageable representation, typically in the form of an inverted index or a similar structure. This representation can be viewed as a row-major sparse matrix where the rows correspond to terms and the columns to documents (that is, a term-document matrix). Inverted indexing starts by parsing each document into a bag of words. Parsing consists of a sequence of simple processing steps: tokenization, stemming, and removal of stop words. A recent algorithm employs both CPUs and GPUs in a distributed fashion to achieve a throughput of 262 MB/s on the ClueWeb09 dataset [Wei & JaJa, 2011]. The strategy includes a pipelined workflow that produces parallel parsed streams that are consumed at the same rate by parallel indexers; a hybrid trie and B-tree dictionary data structure in which the trie is represented by a table for fast look-up and each B-tree node contains string caches (this is particularly useful for indexing on the GPU); the allocation of parsed streams with frequent terms to CPU threads and the rest to GPU threads so as to match the throughput of parsed streams since the different hardware excels at different workloads; and an optimized CUDA indexer implementation that ensures coalesced memory access and effective use of shared memory. Recognizing the complexity of an efficient, GPU-assisted inverted indexing, Ding et al. (2009) found that the implementation of a complete search engine on a GPU does not appear to be realistic, but accelerating query processing is a viable option. Focusing on maximizing throughput on a single machine, the authors worked on decompressing inverted indexes, Boolean set operations, and finding the top-k matching documents for a query.

Topic modelling builds on the inverted index and, given the extremely high computational demand, it has been widely explored as an application domain for GPUs [Lahabar & Narayanan, 2009; Cavanagh et al., 2009; Masada et al., 2009; Yan et al., 2009; Byna et al., 2010]. We add a visual clustering method on top of topic modelling as the final step of our workflow, self-organizing maps, which are known for their enormous computational needs. Visual inspection of data is crucial to gain an intuition of the underlying structures though. As data often lies in a high-dimensional space, we use embedding techniques to reduce the number of dimensions to just two or three. We can also see this step as a unifying approach across various data mining workflows that includes text, images, or other media.

Methods that rely on eigenvalue decomposition, such as multidimensional scaling [Cox & Cox, 1994], achieve a global optimum for such an embedding: the global topology of the space will be preserved. Often the data points lie on a high-dimensional manifold that is curved and nonlinear. These structures are difficult to find with eigenvalue decomposition. Manifold learning generalizes embedding further, assuming that data in a high-dimensional space aligns to a manifold in a much lower dimensional space. For example, the Isomap algorithm finds a globally optimal solution for an underlying nonlinear manifold and it builds on multidimensional scaling [Tenenbaum et al., 2000]. Isomap, however, fails to find non-convex embeddings [Weinberger et al., 2004].

Nonconvex structures are one strong motivation to look at solutions that are not globally optimal, but preserve the local topology of the space instead. Self-organizing maps (SOMs) are a widespread visualization tool that embeds high-dimensional data on a two-dimensional surface - typically a section of a plane or a torus - while preserving the local topological layout of the original data [Kohonen, 2001].

Tool Design

We developed a common computational core that is highly efficient with the available resources. This common core is used as a command-line interface to enable batch processing on cluster resources, but the same core is used as the computational back-end for popular environments in data analysis. The tool, named Somoclu, has the following improvements over other implementations:

- It is highly efficient in single-node multicore execution;
- Large emergent maps are feasible;
- A kernel for sparse data is introduced to facilitate text mining applications;
- Training time is reduced by graphics processing units when available;
- It improves the efficiency of distributing the workload across multiple nodes when run on a cluster;
- An extensive command-line interface is available for batch processing;
- Python, R and MATLAB interfaces facilitate interactive processing;
- Compatibility with Databionic ESOM Tools ensures easy visualization.

The dense CPU kernel is a straightforward implementation of the batch formulation of the SOM algorithm, and it resembles the one implemented by Sul and Tovchigrechko (2011). For an overview of the parallel organization, see Fig. 4-15. The MapReduce calls were replaced by MPI functions, leading to a more streamlined code. Furthermore, single-node parallelism on multicore CPUs is redesigned to use OpenMP instead of MPI. This strategy avoids duplicating the code book: each MPI process must have a copy of the code book, but OpenMP threads can work on the same copy. The simplification leads to a minimum 50% reduction in memory even when only two threads are used.

A performance bottleneck in the original implementation was the accumulation of local weights into a new global codebook by one single process on the master node. This is parallelized by an OpenMP directive. Furthermore, the influence radius $\delta(t)$ of a best matching node is thresholded, which translates to speed improvements without compromising the quality of the trained map.

The GPU variant is more complex compared to the CPU kernel. The complexity stems from the way the distance function is evaluated between the nodes of the SOM and the training data. To maximize parallelism, a matrix of the distances between every data instance and the nodes of the SOM (Gram matrix) is calculated. A naıve approach would be to extend an efficient matrix multiplication algorithm, replacing the dot product by the distance function. Opting for a Euclidean distance, it is possible to derive an alternative formulation of calculating the Gram matrix using linear algebra operations [Li et al., 2010]. Benchmarking the two approaches, we found that the latter approach is a magnitude faster on the GPU, mainly due to a more favourable memory access pattern.

We implemented the GPU kernel with Thrust, a C++ template library for CUDA, which has high-performance primitives, avoiding the need to manually tune individual GPU calls. Compared to the implementation by Wittek and Darányi (2012), the device memory use of the GPU code is reduced to approximately one-third, and the new implementation completely avoids costly matrix transposing operations.

Further complexity arises from the disparity between the number of GPUs and the number of cores in a computer. The CPU kernel achieves maximum speed by running an MPI process on each available core, resulting in a far lower number of data instances per core and a speedier local update of the weights. For instance, if there are eight cores and two GPUs, than each GPU has four times more data to process and its corresponding MPI thread would have four times more data to update the local weights. While the GPU handles the load efficiently, it would be highly inefficient to use a single thread to update the local weights. We thus hybridized the kernel and rely on OpenMP to parallelize the weight update. The GPU implementation runs as many MPI processes on a node as there are GPUs, and uses all CPU cores in the weight update.

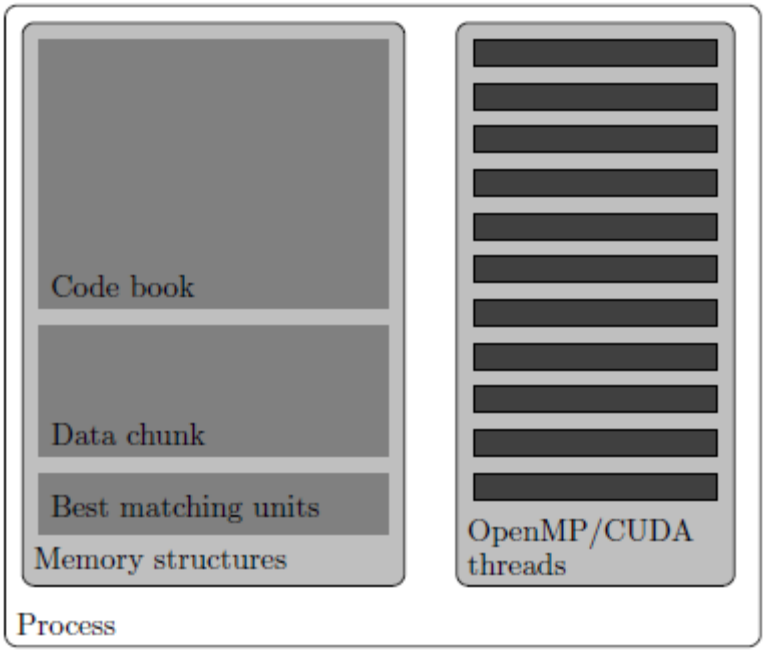


Fig. 4-15. Overview of the parallel organization of the batch training. The global code book is replicated in each process, therefore it is more efficient to use OpenMP-based parallelization than to rely on MPI on multicore CPUs. If a GPU is available, it will overtake most data parallel operations from the CPU, replacing OpenMP threads.

The sparse kernel is a straightforward extension of the dense CPU kernel, and its main virtue is the reduced memory use. A vector space coming from a text processing pipeline typically contains 1–5 % nonzero elements, leading to a 20x-100x reduction in memory use when using a sparse representation. This kernel does not have a GPU implementation, as the irregular access patterns that are inevitable with sparse data structures are not efficient on streaming architectures.

Experimental Results

To ensure replicability of the experimental results, we benchmarked with publicly available cluster GPU instances provided by Amazon Web Services. The instance type was cg1.4xlarge[13], equipped with 22 GB of memory, two Intel Xeon X5570 quad-core CPUs, and two NVIDIA Tesla M2050 GPUs, running Ubuntu 12.04.

[13] https://aws.amazon.com/ec2/instance-types/

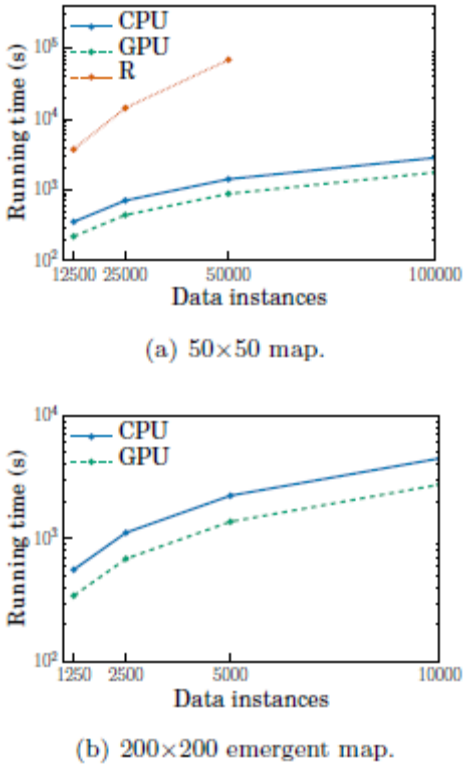


Fig. 4-16. Training time on a single node with CPU and GPU kernels and the R package kohonen. The time axis is logarithmic. The data instances had 1,000 dimensions.

The most direct comparison should be with the implementations by Sul and Tovchigrechko (2011) and Wittek and Darányi (2012). Unfortunately the former was not able to handle the matrix sizes we were benchmarking with. The implementation by Wittek and Darányi (2012) is sensitive to the size of map, and it did not scale to the map sizes benchmarked here, and thus we left it out from the comparison. To compare with single-core performance, we included the R package kohonen [Wehrens & Buydens, 2007]. The number of data instances ranged from 12,500 to 100,000, the number of dimensions was fixed at 1,000 for a regular 50 × 50 self-organizing map. The data elements were randomly generated, as we were interested in scalability alone. We also tested an emergent map of 200 × 200 nodes, with the number of training instances ranging from 1,250 to 10,000. This large map size with the largest data matrix filled the memory of a single GPU, hence giving an upper limit to single-node experiments. Emergent maps in the package kohonen are not possible, as the map is initialized with a sample from the data instances. If more the map has more nodes than data instances, kohonen exits with an error message. The map size did not affect the relative speed of the different kernels (Fig. 4-16).

Compared to the R package, even the CPU version is at least ten times faster. The difference increases with the data size, indicating serious overhead problems in the R implementation.

The GPU results show at least a two-times speedup over the CPU version. This is less than expected, but this result considers a single GPU in the dual configuration, and the CPU is also a higher end model. These results, nevertheless, show that the Thrust template library is not efficient on two-dimensional data structures.

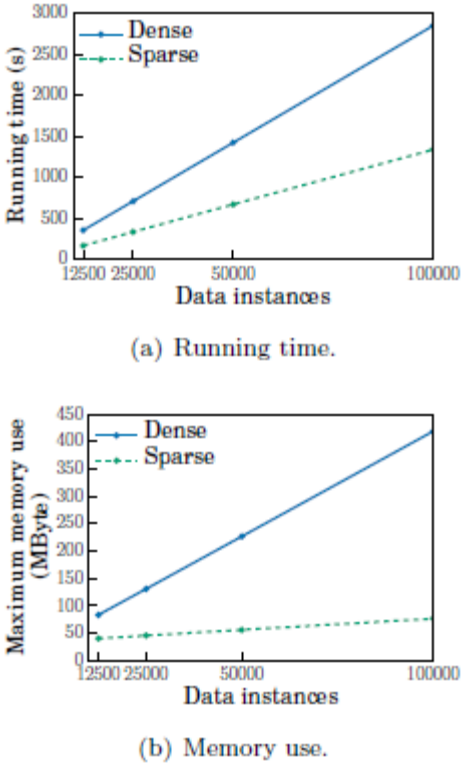


Fig. 4-17. Training time on a single node with dense and sparse kernels. The data instances had 1,000 dimensions, with five per cent of the elements being nonzero.

Comparing the sparse and dense kernels on a 50 × 50 map, we benchmarked with random data instances of 1,000 dimensions that contained five per cent of nonzero elements (Fig. 4-17). Execution time was about two times faster with the sparse kernel. The reduction in memory use was far more dramatic, the sparse kernel using only twenty per cent of the memory of the dense one with 100,000 instances. Naturally, the difference with emergent maps would be less apparent, as the code book is always stored in a dense format.

Measuring the memory overhead of the interfaces compared to the native version, we are not surprised to see that the Python variant is the closest (Fig. 4-18). As the data structures are not duplicated in this interface, it is still counter-intuitive that the gap increases with larger data sets. The R and MATLAB versions have predictably larger and growing gaps compared to the command-line interface, as they both must duplicate all data structures. Apart from the time spent on duplicating the data structures in the R and MATLAB versions, the computational overhead is negligible in all interfaces.

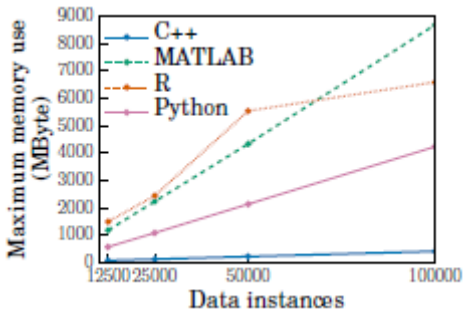


Fig. 4-18. Memory overhead of the Python, R, and MATLAB interfaces compared to the command-line version (indicated as C++).

Using 100,000 instances and a map of 50 × 50 nodes, the calculations scale in a linear fashion (Fig. 4-19). This was expected, as there is little communication between nodes, apart from the weight updates. As calculations cannot overlap with calculations, we did not benchmark the GPU kernel separately, as its scaling is identical to that of the CPU kernel.

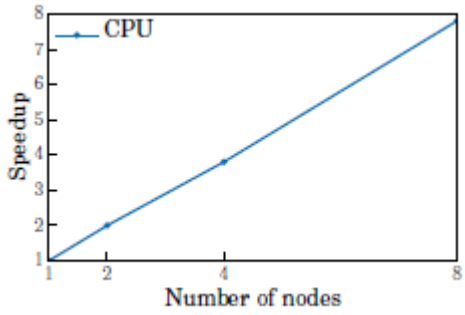


Fig. 4-19. Speedup on multiple nodes with CPU kernel compared to a single node. The data instances had 1,000 dimensions.

Feasibility Pilot on Two Datasets of Semantic Content

Having shown above that Somoclu is significantly faster and scalable than other approaches and available tools in this development area, for reasons of consistency in presenting Somoclu-related results together, we now turn to its combined testing where scalability meets the automatic extraction and classification of collection content based on features extracted from text-based documents. The rest of the adopted text analysis methodologies are presented in Section 4.2 below.

Scalability aspects:

To test the scalability of Somoclu on the field model, we repeated the experiment based on Stanford's Amazon book reviews data set as a collection of digital objects [McAuley & Leskovec, 2013], which is publicly available as part of the University's SNAP project. The data set spanned a period of 18 years and included approximately 12.8M book reviews up to March 2013. Every item in the data set included product and user information, ratings, as well as a plain text content description. We split the corpus in three periods, each containing close to 4.3M objects. The key characteristics are summarised in Table 4-2.

Table 4-2. Key statistics of the temporal split of the corpus.

Period 1	Period 2	Period 3
Until 30 Jan 2003	Until 03 Aug 2008	Until 04 Mar 2013
45162 terms	49400 terms	50672 terms

The results were evaluated for their semantic consistency based on their statistical significance [Wittek et. al, 2015a]. More research will be necessary to work out a comprehensive evaluation methodology to interpret the interplay of position and direction vectors. As the latter indicate emergent changes in the field, the dislocation of actual semantic content vs. potential displacements will have to be addressed in terms of a dynamic theory of word semantics. This point in the direction of concept drifts and topic shifts as related research areas.

Applicability to artworks metadata:

For a dry run on artworks and media metadata, we tested Somoclu on the Tate art collection and archive holdings dataset publicly available for research. The dataset contains approximately 70,000 records with the following metadata elements for each artwork and archive item:

- Artist(s)

- Title
- Date created
- Reference number / Accession number
- Medium description
- Web address (URL) of page in Art & artists section of the Tate website
- Subject index terms
- Image web address (URL)
- Credit line
- Movements

To get a first impression about the evolution of artistic technology between 1823-2013, we generated a 69000 x 1023 binary (presence-absence) matrix of artworks indexed by the medium used for self-expression. The results (Fig. 4-20) indicate that in a next phase of research, index terms from the hierarchical Tate subject index will be a suitable input to map trends of semantic content evolution.

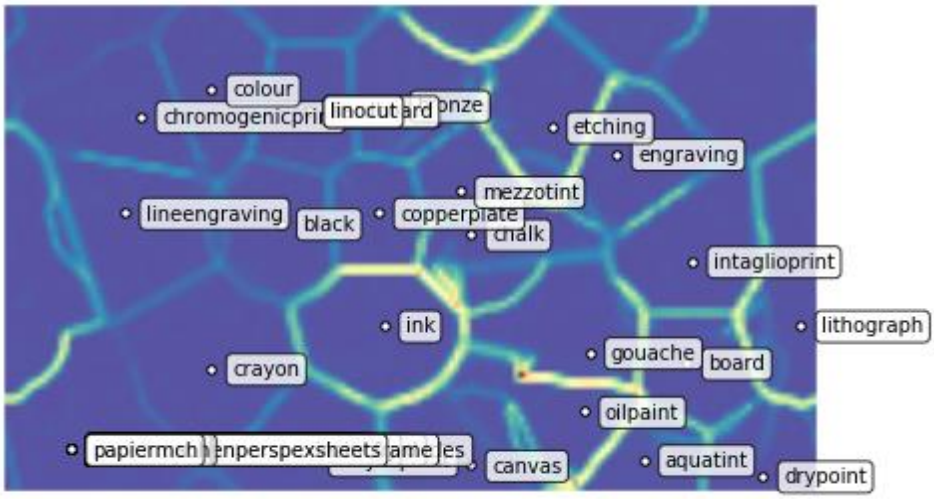


Fig. 4-20. A small, 90 x 150 nodes toroid map separating artistic media of self-expression over 190 years with boundaries indicating established vs. incoming trends. Partly due to the impact of the Turner bequest, old techniques are dominant.

Exploring the Mathematical Basis of Machine Learning Results

To round off our considerations about tools to address the question whether quantum likeness of data could be an issue for LTDP, as said above, testing quantum-likeness needs specific mathematical software. A hierarchy of semidefinite programming (SDP) relaxations approximates the global optimum of polynomial optimization problems of noncommuting variables. Generating the relaxation, however, is a computationally demanding task, and only problems of commuting variables have efficient generators. As a prelude to D4.5 and to meet its specific tool needs, we developed an implementation for problems of noncommuting problems that creates the relaxation to be solved by SDPA – a high-performance solver that runs in a distributed environment, called Ncpol2spda ("Sparse Semidefinite Programming Relaxations for Polynomial Optimization Problems of Noncommuting Variables"). We further exploit the inherent sparsity of optimization problems in quantum physics to reduce the complexity of resulting relaxation. Constrained problems with a relaxation of order two may contain up to a hundred variables. The implementation is available in C++ and Python. The tool helps solve problems such as finding the ground state energy or testing quantum correlations [Wittek, 2015].

The importance of this software lies in the fact that high-performance supporting libraries for solving SDPs do exist: they are technically capable of scaling for large problems. However, the so far missing

link has been a tool to generate the SDP relaxation of a given order from a polynomial optimization problem of noncommuting variables; this was developed and made available online under an open source license. The tool is able to generate the SDP relaxation of a hundred variables in about thirteen hours, although generating a sparser relaxation takes much longer. Improvements in the underlying symbolic libraries could improve execution time, especially by adding more efficient hashing functions, ensuring thread safety, and addressing the efficiency of substring replacement. A key application field would be finding the ground state energy of an arbitrary Hamiltonian. It remains to be seen how this approach would compare to other methods in execution speed and accuracy.

4.2. Analysis of Text-based Content

This subsection discusses the process and tools for analysing source text documents relevant to the two case studies and using the extracted information for populating the developed domain ontologies with instances.

4.2.1. Science Case Study

Documents Relevant to the Science Case Study

In the Science case study, the following kinds of documents are relevant for preservation and from which information extracted is valuable[14].

Engineering Documentation

The first considered type of documentation are those documents (923 in total) that were created in the development phase of the SOLAR experiment, typically before the operational activities of the experiment have started. These documents include specifications and design documents, safety and acceptance data packages, test plans and reports and Interface Control Documents. From these documents, metadata such as the title, author, creation date and so on can be extracted and this information can be used for relating the different documents to each other, but also for relating them to activities and products (data or reports) during the operational phase of the experiment. Often, it is relatively straightforward to extract this data from the documents, provided that a known format and structure is consistently followed. However, considering the (international) scale and complexities of the engineering endeavours involved, this is less than self-evident in reality.

> **An example**: *Consider a 'console log file', produced by an experiment operator, wherein up to minute by minute noteworthy events of the day are manually logged, in free text. Time pressure doesn't always allow the operator to look up and provide sufficient details to refer to a particular engineering document. Instead, the operator might refer to just a title (or a variation thereof). For long-term thorough understanding of the log file, it is paramount that a link can be established and preserved between that log file and the referred engineering document.*

Apart from metadata, also information in the documents themselves can be extracted and used for linking against operational products. However, as most documents in this category are one-off documents, a lot of manual fine tuning and data extraction effort is needed to capture to a certain amount of detail what is in the documents. A careful trade-off is to be made between the required data that is needed from the documents and the effort required. Low effort, high reward extraction

[14] More detailed explanations about these documents can be found in D2.3.1 "Functional Requirements & User Descriptions" [PERICLES D2.3.1, 2014] and in D2.3.2 "Data Survey & Domain Ontologies" [PERICLES D2.3.2, 2015].

methods (such as relying on chapter/section titles or even figure/table captures, or document classification techniques based on term frequencies) are likely to prove the most efficient approach.

> *Continuing the example above, the console log might refer to so called operation modes of parts (eg. the SOLSPEC instrument) of the SOLAR Payload. Title analysis of chapters in the 'SOLAR Payload Operation manual' will allow to directly establish a link between this console log entry and the correct section in the correct document. This will allow faster interpretation of the console log entries at a later stage. In addition, a user needs to learn about the SOLAR payload, it will be valuable to him that he can browse from the semantic model itself to the relevant sections of the manual. For example, the user who browses to the SOLAR payload and from there to the SOLSPEC instruments could then see that there are operating modes to SOLSPEC and browse further to them.*

Operation Documentation

During the active operational phase of a space science experiment, an enormous amount of documentation and data is generated and used. Therefore, there are many types of operation documentation that are useful to preserve together with the data.

Rules and Guidelines

The documents in this subcategory (such as Operations and Interface Procedures, Flight Rules and Regulations, Operation Data Files) could be relevant for tracing how and why certain operation actions were taken. They can be useful to better understand the operation logs. Again, the metadata of these documents and for example titles, subtitles, and structure, can be useful to relate the documents to each other.

> *For example, documents reporting on anomalies might refer to Flight Rules and Payload Regulations. These links are very important to be preserved as without them, it might prove difficult to, in the future, sufficiently understand the rationale behind measures taken to avoid reoccurrences of the anomalies.*

Actual Operation Documentation

There are many types of documents that discuss the actual operation details – these include the detailed planning of the operation, communications between the operation people and between the operation people and the scientists, and logs of the different events that occur during the operational phase of an experiment.

As before, for all these documents, the respective metadata can be valuable – titles, author(s), dates. For example, the dates of the minutes of meetings, the daily operation reports and the console logs can be useful to understand which of those are related to certain command schedule[15].

However, in some cases, extracting more detailed information using natural language processing techniques from the text itself might be interesting. For example, finding exactly when certain command schedule was running during a shift (in other words, in which Activity, reported on in a specific Console Log) can be useful as it allows to relate incoming telemetry and scientific data to the commands that triggered them.

It should be noted that often it is not trivial to parse and understand operational documentation, as, while in some cases the same formulation, conventions and structures are used, often the concrete application of these is imperfect: spelling errors, grammar errors, shortcuts made that can only in the short term be understood by experienced operators, omissions because of prioritisation, inconsistent interpretation of the (evolving) guidelines and so on.

[15] Command schedules are scripts listing time-tagged commands to the instrument of SOLAR Platform.

Scientific Results

This category includes all experiment outputs that are generated after the operational phase. For example, incoming raw scientific data is typically calibrated and processed into scientific data that can be interpreted and used for scientific research. The category also includes scientific papers and reports. Even though being able to link these scientific outputs to earlier outputs and activities is valuable (e.g. being able to backtrace claims made in a paper to its ultimate source, the execution of specific commands on the experiment hardware, triggered by the occurrence of a certain event), these links are currently not further explicitly considered in the scope of this project as the use case partner B.USOC operates the SOLAR experiment, but does not disseminate/process its raw results into scientific outputs and therefore has no direct access to the material and knowledge needed.

Data Stores

The various types of documents and data described in this section so far are in practice stored in or generated using multiple data stores of various types. The data stores, located at the premises of case study partner B.USOC, that were considered most relevant for the case study are included in Table 4-3.

Table 4-3. Data stores relevant to the Science case study.

Data Store	Type of data store	Description
Alfresco and Dropbox	SMB[16] derived Document stores	Both document stores various operations and design related documents such as user manuals, console logs, operation reports, reference material. Alfresco (a CIFS based content management system) shall supersede Dropbox (a samba share).
Predictor	Operations software with relational database backend	A central access point used by experiment operators. Contains shift information, activity information, data stream information, event information and so on.
YAMCS	Operations software with column-oriented DBMS backend	YAMCS is the main telemetry and data archive used for SOLAR data at B.USOC.

In addition to these data stores, a significant part of the domain model is populated completely manually. As an example, the simple fact that the SOLAR experiment uses 3 different instruments (SOLSPEC, SOVIM and SolACES) is something that -even though obviously this knowledge is captured in documents- is added manually into the domain model. Trying to automatically extract this kind of one-off knowledge is inefficient at best.

Population

Connectors

These five data sources (four data stores and the manual population) are used to populate the domain model that is based on the domain ontology as described in D2.3.2 [PERICLES D2.3.2, 2015].

[16] https://en.wikipedia.org/wiki/Server_Message_Block

In the context of the development of the Space Science Portal[17], software components are under development that provide the ingest from data stores' functionality. These components are called **Connectors**.

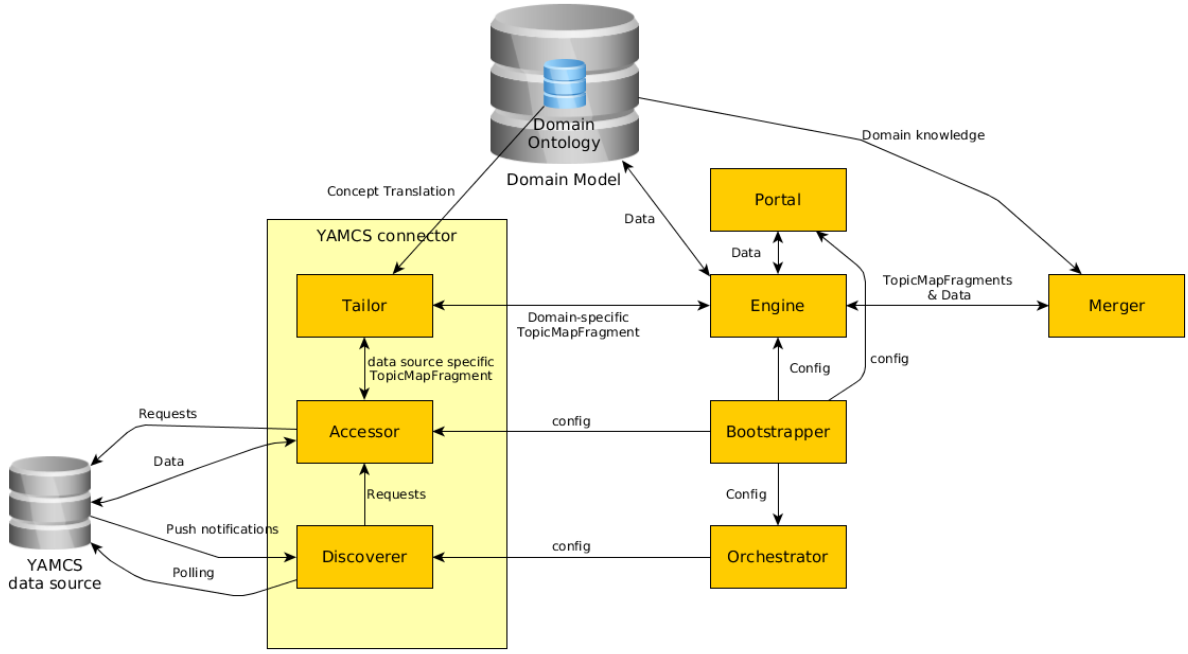


Fig. 4-21. Connectors provide the link between data stores and the domain model.

There is one connector per data source and each connector is composed of a few subcomponents:

- An *Accessor* is responsible for opening and managing the connections to the data store, and provides the mechanisms to request and access the data and, if applicable, to write data into the data store. This subcomponent is data store specific.
- A *Discoverer* interface with the *Accessor.* This subcomponent ensures that the domain model is in sync with its data store. Depending on the data store, this synchronisation can happen in real time meaning that any new data arriving in the data store is immediately reflected into the domain model. Alternatively, the synchronisation can be user-triggered or scheduled if real time synchronisation is to be avoided (e.g. when synchronisation requires locking the data store or is computationally expensive).
- A *Tailor* subcomponent provides the translation between the data/documents, as accessed by the *Accessor*, and the domain ontology. This allows to reuse a big part of the Connector and use it in a different context (i.e. with a different ontology), simply by injecting a new Tailor or updating an existing one. Data that has passed through the Tailor is ready to be injected into the Domain Model, using an Engine component that is developed specifically for the Domain Model storage technology that is used.

The connectors connect to the Space Science Portal backend that contains an orchestrator which coordinates how and when connectors should communicate with their respective data store. Data communication between the connectors and the backend happens through *TopicMapFragments*, which are relatively simple data containers that are inspired by the Topic Maps structures but that in principle can be used on any type of backend, if a lightweight suitable adapter is developed.

[17] An early version of the Space Science Portal was presented in the first review. The Portal is a web application which takes advantage of the semantic model in order to allow the user to browse between the digital resources and information about them.

TopicMapFragments generated by an Accessor for a specific data source are specific to that data source. The Tailor makes them domain specific. A *Merger* component takes inputs from several Connectors and is able to make associations between them.

Data Extraction

At the start of the project it was assumed that, given the available data and documents that most metadata and data that can be found in the various (engineering and operations) documents only by parsing those documents. However, only during the project, after discussions with B.USOC and an analysis of the data store software used (including the schema's of those data stores), it became clear that in fact the most interesting information can be found as metadata in the data stores themselves and are directly accessible. For nearly all documents, author, creation date, subject and type of document can be easily retrieved from the data stores. Also very valuable to learn was that much more operations-related information is stored in the Predictor software, information that was originally assumed to be created manually. For example: Daily Operations Reports (DOR) of SOLAR are created daily and contain an overview of everything that happened with respect to the SOLAR experiment during that day: shift, sun visibility, activities, events, anomalies and so on. It turns out that this whole DOR is in fact auto-generated using the Predictor software. Consequently, it makes much more sense to configure our Predictor connector so that its accessor allows us to extract this data directly from the Predictor backend, as opposed to using NLP techniques to try to extract this information from the reports themselves. Depending on the extracted data, the connector might need to process the data before it can be inserted in the domain model. In particular, some multi-element data is stored in the Predictor software as space- or comma-separated strings of text, some data is only accessible by applying natural language processing on non-trivial strings that are stored in Predictor and some data needs to be interrelated using multiple tables before the necessary associations can be created in the domain model.

Note however that even though this simplifies data extraction a lot, it is not a one fix for all. Indeed, certain operations related aspects that are found in Predictor are entered in a free-form format that still requires humans to understand it. For example, the execution of a Command Schedule is logged as an entry in a table in Predictor. The detailed off-nominal timings of parts of that Command Schedule, are entered manually by an operator, in a field of that entry. Fully understanding the execution of the Command Schedule, which is critical to be able to relate the data to it, still requires parsing these manually entered fields.

Concrete Example

The following is an example of a typical flow of how the connectors are deployed and used to automatically populate the space science domain model, based on the data stores that are operational at the B.USOC premises.

1. Bootstrapper launches Dropbox connector and configures its Accessor to connect to Dropbox data store
2. Bootstrapper triggers Orchestrator which triggers Dropbox Discoverer to do a daily check of Dropbox repository to check for new data
3. Dropbox Discoverer polls the Dropbox data store and finds a new Daily Operations Report (DOR)
4. Dropbox Discoverer asks Dropbox Accessor to retrieve the metadata from the Dropbox data store
    a. Dropbox Accessor retrieves metadata
    b. Dropbox Accessor creates a TopicMapFragment with the metadata
        - *Topic ID: DOR_BOPS_2015_23.pdf*
        - *Topic Occurrences:*

        - *Filename: DOR_BOPS_2015_23.pdf*
        - *Generation Time: 2015-01-24, 00:10*
        - *Source URI: Dropbox://BUSOC/SOLAR/OPS/2015/Jan/DOR_BOPS_2015_23.pdf*
        - *Author: Nadia This*
5. Dropbox Tailor tailors the TopicMapFragment by using concepts from the Domain Ontology. E.g. the tailor can add/change the following
    - *Topic Name: DOR_BOPs_2015_23*
    - *Topic Name Variant: Daily Operations Report, 2015, Jan, 23*
    - Topic Type: Daily Operations Report
6. Dropbox Tailor sends the modified TopicMapFragment to the Engine that sends it to the Merger component. The Merger component uses the domain model (that has previously used the Manual Connector to populate the domain model with concepts such as persons, types of shifts, solar experiments and instruments, types of data sources). The Merger updates the TopicMapFragment to add more associations. New TopicMapFragment:
    - *Topic ID: DOR_BOPS_2015_23.pdf*
    - *Topic Name: DOR_BOPs_2015_23*
    - *Topic Name Variant: Daily Operations Report, 2015, Jan, 23*
    - *Topic Type: Daily Operations Report*
    - *Topic Occurrences:*
        - *Filename: DOR_BOPS_2015_23.pdf*
        - *Generation Time: 2015-01-24, 00:10*
        - *Source URI: Dropbox://BUSOC/SOLAR/OPS/2015/Jan/DOR_BOPS_2015_23.pdf*
    - *Associations:*
        - *Association ID: association1*
        - *Association Type: Authored By*
        - *Role 1: Author: Nadia This*
        - *Role 2: Document: DOR_BOPS_2015_23.pdf*
7. The resulting TopicMapFragment is then sent back to the Engine that merges it with the domain model.

4.2.2. Art & Media Case Study

Extracting Information from Social Media Sources

This area of research concentrates on evaluating the utility of social media sources to supplement traditional information relating to artworks. In particular, the major aspect of this work is to determine what types of information can be extracted from social media. Looking in particular at data collected from microblogging platforms Twitter and Tumblr and the photo-sharing website Flickr, we describe the aspects of data in relation to Tate that we can harvest and potential uses of this data in supplementing existing sources.

Data Collection Methodology

Each platform makes available a variety of APIs (application programming interfaces):

- **Twitter** provides a rate-limited search interface with a fifteen minute time limitation: this retrieves only a proportion of search matches. This is authenticated via OAuth and returns query responses in JSON, interpretable through compatible libraries such as Python's simplejson. Alternative APIs made available by Twitter include the high-volume Streaming API, Firehose and the comprehensive but expensive Gnip service for retrieval of historical

Twitter data, none of which are proportionate for the present study. The search interface may be queried using various search modes: here we apply keywords relating to Tate as a trade-off between search precision and recall. It is likely that using a further geographically bound search covering the areas of each Tate gallery would identify further tweets of relevance to material contained within Tate, but estimates suggest that only 1-3 percent [Morstatter et al., 2013] of tweets are geocoded, meaning that the benefits of this approach are limited.

- **Tumblr** provides a somewhat analogous OAuth - authenticated search API: in this instance there is no limit to the proportion of data returned and no temporal restriction on accessibility of material. With the exception of material later deleted or removed by authors, all historical material on Tumblr remains accessible. In previous studies, however, there are signs that such deletion is not uncommon.
- **Flickr** does provide an authenticated API, but this is not geared towards large-scale search or browse of the material contained on the service. Instead, this material can be searched using Flickr RSS (syndication) functionality, which allows for various search mechanisms such as the setting of specific temporal windows. Although the RSS is limited in the number of items it can return, these mechanisms effectively permit the user to work around the limitations by careful query formulation.

Whilst the API differed in each case, we store all returned content in separate SQLite database files.

Filtering Input for Relevant Content

Initial filtering of each data source is minimal: that is, we search for 'tate' rather than 'tate gallery' in the case of Twitter, even though this also returns terms such as 'state' or 'estate'. This is done to increase recall (the fraction of relevant instances retrieved, see [Jardine & van Rijsbergen, 1971]) at the earliest stage, allowing an investigative approach to analysis and limiting bias at the earliest stages. Developing a strategy for optimising filtering is non-trivial [Abel et al., 2012] and consequently our approach is iterative in nature. For the purposes of the present document our intent is exploratory: hence, we interest ourselves more in characterising what is present than in optimising a strategy for returning a particular profile of material.

In all three cases, the substring 'tate' was used as an initial search term, alongside (in the case of Tumblr) search for the lengthier expression 'tate gallery'; this is done to pick up specific formulations that may not otherwise be retrieved. Due to Tumblr's design, it is convenient to search for specific tags rather than solely for terms, and therefore a snowball methodology is used to spiral outwards from initial hits to other posts or blogs that may be of relevance ([Biernacki & Waldorf, 1981; Atkinson & Flint, 2001]).

Use of the substring 'tate' in case-insensitive search system captures hash tagged tweets or tagged posts, as well as mentions of the term itself. A proportion of false positive terms, which we will discuss briefly in our analysis, are also retrieved.

For the purposes of the present evaluation we apply strict filtering rules in order to limit material returned to material containing either the string 'tate' with appropriate word boundaries, or material containing Tate's hostname.

Of the original 222,356 tweets collected between 20 Feb 2015 and 20 Nov 2015, the sample evaluated consisted of 22,000 tweets.

For the Tumblr data, a search was completed for posts of any age containing the term 'Tate'. Of the original 70,000 posts from 01-Feb-2005 to 24-03-2015, 3,093 were examined in this present process. This implies that the snowball sampling method applied on Tumblr provides a far lower rate of recall than the keyword search strategies applied on the other two sites; this is unsurprising since snowball sampling depends on serendipity and tends to give a low rate of precision in a highly connected and

diffuse network: most of the blogs identified on Tumblr are not solely about Tate, or even art in general, and many of them also link to material that has no relation to either.

In the case of Flickr, 144,489 photographs are retrieved; their dates posted ranged from 1906 to the present day, and dates uploaded ranged from 2006 to the time at which the sample was taken (Feb 2015).

A random sample of 10 cases from each of the resulting data sets was qualitatively evaluated by hand in order to better understand the information available. This is a valuable first step as it gives us a rough metric regarding the quality of the data for the current purposes and what quantitative methods might be appropriate for evaluation of the whole data set.

Qualitative Initial Survey

A random sample of 10 cases of posts/tweets from each of the three data sets was collected using the 'random()' query in sqlite. As noted previously, this is in order to gain a sense of the social media content mentioning Tate, in order to scope out potential applications of the data. In the following, we describe the patterns emerging from this data.

```
Henry Moore Foundation Aug 2010|photo|family sculpture brick green
london art field grass yellow stone modern barn studio print square
design etching gallery arch treasure sheep tate group plaster
exhibition foundation canvas plastic relief sundial henry moore
cast picasso bourne much create reclining draw portfolio piece
shape perry oval irina auden locking saatchi maquette hadham
durrell hoglands|2010-08-22 13:03:33|nikon_13|51.837719|16|Figure
In A Shelter (1983)
```

Fig. 4-22. A Flickr image of an artwork contained within the Tate collection.

From the 51,515 Flickr posts, the 10 which we examined displayed a range of characteristics: two were centred upon presenting images of artworks (presumably) contained within the Tate galleries' collection (see Fig. 4-22), with one concerned with an image of the inside of a gallery at Tate Britain (focussing more on the room rather than the number of art objects visible on display) and another shows the inside of the Tate Modern turbine hall with no artworks visible, with both of these latter two images perhaps indicating a greater concern with the infrastructure and ambiance rather than the artworks on display.

```
The Millennium Bridge|photo|london millenniumbridge
tatemodern southwark cityoflondon city thames bridge|2005-
06-08 13:38:21|ickoonite|0|0|The Millennium Bridge and its
shadow in early June.
```

Fig. 4-23. A Flickr image of a Tate Modern building.

Three more image posts are focused in the iconic nature of the Tate buildings: two of the Tate Modern building shown against the London skyline (see Fig. 4-23), with a third showing the Tate Liverpool building. The final two images were apparently irrelevant pictures of London and Liverpool, which may have been tagged mentioning 'tate', or, more probably, either contain a building or other visual reference of relevance to Tate or relate to an activity featuring Tate. That is, if an individual is visiting a gallery they may tag all the photographs related to that activity in this way, even though not

all photos relate obviously to that entity. To summarise, two of the 10 images directly relate to artworks, with six more related to the architecture and ambiance of the buildings, with the final two apparently irrelevant or of limited direct relevance to Tate.

Of the 22,292 tweets collected, 450 were identified as being directly relevant to the Tate galleries: six tweets directly referenced an image relating to an artwork (e.g. see Fig. 4-24(A)), with five of these purely describing the object (in some cases text is simply copied from the Tate website or catalogue), and one of these including a reaction to an artwork. Two tweets referenced news stories mentioning Tate, one of which was a promotion of a new exhibition (e.g. see Fig. 4-24(B)) and the other was a negative news story about Tate's unwillingness to loan an artwork to another gallery. Of the two remaining tweets, one mentioned the Tate gallery in passing with reference to another art work, and the final tweet made no obvious reference to Tate. Therefore, in summary 6 of the 10 tweets directly related to art objects, with the two news stories more relevant at an organisational level.

```
RT @RobertFehr3: Marc Chagall, 'Bouquet
with Flying Lovers', (c. 1934-7), oil on
canvas, Tate Gallery, London
http://t.co/VIzfOU5J4r
```

A

```
RT @standardnews: These amazing
slides are coming to the South Bank!
http://t.co/yYnmBnTJvq
http://t.co/IAwg8QeQc0
```

B

Fig. 4-24. Twitter excerpts.

In the case of Tumblr, 3,093 posts were deemed relevant, and the 10 analysed related to Tate as follows: four were images of art objects (although only one was a personal image containing an individual's opinion (see Fig. 4-25(A)), the rest are images and descriptions copied from existing sources relating to the Tate's collection; for example see Fig. 4-25(B)). Two posts related to publicity (press releases, interviews, news items, on Tate's site) published by Tate relating to exhibitions, and two further posts mention Tate with reference to broader discussions about the art world. One post relates to a personal story of a visit to the Tate, and one more post is irrelevant (it mentions someone named 'Tate'). By way of summarising these results, four posts relate directly to art objects, with five more generally relating to the Tate organisation.

```
Roman Ondak's
Measuring the Universe
at Tate St Ives stem
from the idea of
parents measuring and
marking the height of
their children on the
door frame. Love the
concept. Art with a
temporary existence.
```

A

```
Mountain Lake demonstrates Dalí's use of the multiple
image: the lake can simultaneously be seen as a fish. By
such doubling he sought to challenge rationality. The
painting combines personal and public references. His
parents visited this lake after the death of their first
child, also called Salvador. Dalí seems to have been
haunted by the death of his namesake brother whom he never
knew. The disconnected telephone brings the image into the
present by alluding to negotiations between Neville
Chamberlain, the British Prime Minister, and Hitler over
the German annexation of the Sudetenland in September
1938.
```

B

Fig. 4-25. Tumblr excerpts.

In terms of making sense of the social media data types and possibilities for incorporating such data into a gallery organisation, such as Tate, we note two main categories: firstly posts relating to specific artworks (and exhibitions) and posts relating to the organisation. There is of course an additional category of less directly relevant posts (4/30 posts), with this perhaps lower than may be expected in social media data (although we expect that this to some extent is due to the fairly rigorous filtering employed in processing/selecting the posts). The first category of posts directly relating to specific objects does not provide new information about the objects (often this is copied from existing sources), however in some cases there is evidence of evaluation and opinion by the author from their text. Even without such opinion, this data would be potentially useful for incorporating into altmetric measures of artworks and their popularity or visitor access (e.g., how many times an object has been posted, re-tweeted, linked to, etc. cf. e.g. [Priem et al. 2010]). We expect similar metrics can be applied to exhibitions as we propose for individual objects, however we note that this data may be skewed by many of the posts being produced as publicity by Tate themselves, or coverage by e.g. newspapers.

The second main category of posts relate to the organisation (Tate) itself. This category can be further filtered depending upon how central Tate is to the post, or whether as in some cases it is mentioned in passing. Interestingly, we note that in the case of Flickr, the buildings housing Tate collections are often seen as art objects in themselves. Again, altmetrics could be applied to institutional-level information, to give a sense of opinion, access or even relative status. In all cases of social media data provides network data describing relationships, such as from re-tweets, or even patterns of users similarly posting about objects: this could be used in order to build social network analyses relating to artworks, users, and institutions.

In summary, this qualitative evaluation suggests that following the aggressive filtration of our initial sets, we are able to demonstrate an 80-90% precision. However, we have not calculated recall in this instance, and it is likely that recall has suffered due to the aggressive filtering method. In practice, we may wish to find a more optimal balance between the factors of precision and recall in order to improve recall. This would require further evaluation.

Quantitative Analysis

Techniques drawn from text mining may be applied to metadata such as that discussed here to extract broad trends and provide an overview analysis of datasets. Here we describe findings from initial analysis of the three sources.

Method

Basic corpus linguistic analysis is carried out using a number of R libraries, notably R's 'tm' (Text Mining) and hierarchical cluster analysis functionalities. Each set is initially normalised by removal of any extant markup such as HTML and link-based navigation, as well as any features specific to particular platforms such as @username references. The material is then processed into lowercase and punctuation and Unicode representations are removed. English-language stopwords are also removed. At this stage we do not include language detection in the pipeline, although it is suggested that identification of the languages used would at a later stage be a relevant task. This is particularly true since the amount of material available relating to cultural resources in languages other than that of the country of the gallery is generally very limited, and hence it is possible that material in other languages would be particularly valuable to a variety of user communities. We also do not stem at this stage. Each dataset is then processed into a vector space representation [Sahlgren, 2006].

Initial observations

We observe that text length varies from a mean of 95 characters on Twitter (standard deviation 26.2, max 156 chars), to a mean of 168 chars on Tumblr (standard deviation 249.7, max 4331), to a mean

of 566 characters on Flickr (standard deviation 1475, max 65720 characters). This suggests that texts on the latter two platforms are applied for a variety of purposes i.e. in some cases they are simply titles or very brief descriptive strings: in others, lengthy texts are provided.

Exploring term association

The use of statistical measures of similarity, such as distance evaluation within a term-document matrix, is well established within corpus linguistics [Jiang & Conrath, 1997], and a number of related techniques have applied such an approach [Lund et al., 1995; Landauer & Dumais, 1997; Turney, 2002]. The fundamental assumption of this approach is that a word's meaning is represented by "the company it keeps" in a large corpus of text [Firth, 1957]. Therefore, the average degree of physical proximity over a large number of texts of two words is a measure of their semantic proximity, with this encoded within a term-document matrix representation.

Using term-document matrices for each dataset, we query for sample terms in each dataset:

Table 4-4. Term associations for the terms 'art' and 'gallery'. Values in brackets represent term associations.

Term	Tumblr	Twitter	Flickr
art	history (0.58) painting (0.42) contemporary (0.39) english (0.39) british (0.35)	1915 (0.42) exchanges (0.42) international (0.42) 1965 (0.41) cornwall (0.41)	upsidedown (0.87) multidirectional (0.85) dollars (0.84) efforts (0.84) emerson (0.84)
gallery	national (0.33) london (0.22)	tate (0.35) london (0.28) art (0.23) ives (0.21) modern (0.19)	exhibition (0.43) national (0.42) british (0.37) works (0.36) Wikipedia-Tate Britain (0.36)

At first blush Flickr's term proximity results for 'art' may seem improbable. In fact, the initial term ('upside-down') reflects a consistently formatted cluster of Flickr pages relating to Emerson's Upside-Down Art or Masg Art.

It is notable that discussion of Tate St Ives takes a higher profile on Twitter relative to the other sites.

Visualisation of Primary Clusters in Data

For visualisation purposes, infrequently occurring terms are removed. Distances between terms in the term-document matrix are then calculated using Euclidean distance, and the terms are then clustered using Ward hierarchical clustering [Johnson, 1967]. The result is plotted on a dendrogram. This analysis shows the primary (most prominent) clusters identified in each dataset.

The Twitter dendrogram (see Fig. 4-26) shows a large cluster surrounding terms related to Cornwall. Notably, 1915 to 1965 refers to an exhibition taking part at Tate St Ives. The third cluster referring to curator tours appears to relate to Tate activity clusters referred to under the title 'curator tours' (effectively a cluster of 'curator tour activities and responses'). The large cluster on the far right of the Twitter dendrograms contains a very large number of terms, although most of them are too low in volume to be visualised here: this cluster refers to activity relating to visiting or discussion of Tate London. The 'tate gallery' area to the left is a high volume cluster (the height at which each term appears relates to the number of times each term appears), which seems to relate to the many tweets referring to something, in broad and general terms, as present 'is in the Tate Gallery'. If we

looked deeper into that cluster it is likely that many individual artist names would appear therein. This can potentially be interpretable as a form of pure 'citation' cluster.

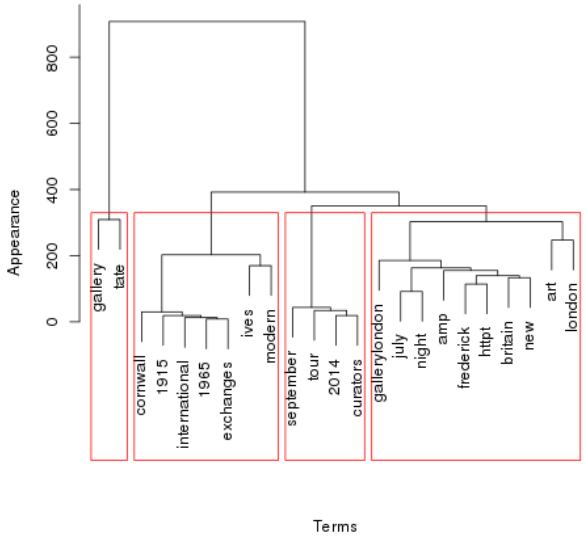


Fig. 4-26. Dendrogram of Twitter terms Spring-Autumn 2015.

So, in summary, from the six months of Tate-related Twitter data, clusters relate partially to geographic factors (Tate St ives), partially to exhibitions (and note that the 1915-1965 area occupies a different branch of the dendrogram to the '[st] ives' term), and partially to citation. Although these findings are interesting as an overview, currently we do not relate them to activities or exhibitions; this would be an interesting avenue of investigation (especially in relation to the altmetrics approaches).

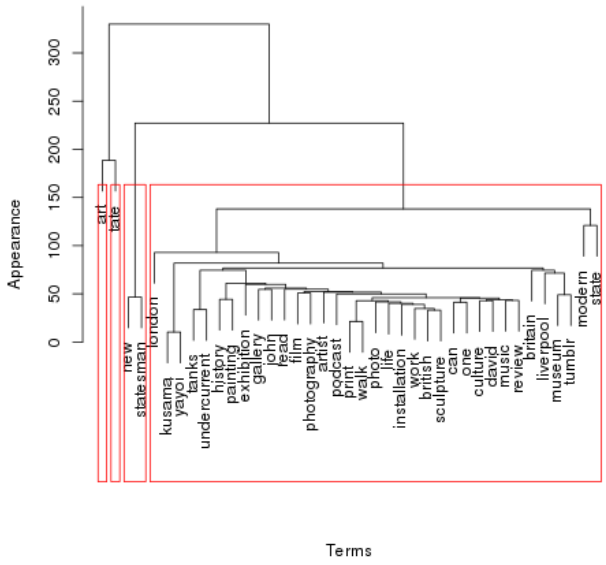


Fig. 4-27. Dendrogram of Tumblr terms relating to Tate.

The Tumblr data (see Fig. 4-27) demonstrates some interesting characteristics. Notably, we observe that the material goes into a great deal more depth; unlike Twitter, Tumblr posts are not limited in terms of length, and therefore some posts are relatively verbose. This may contribute to the impression that these posts are relatively information-rich. We also note that mentions of St Ives occur almost exclusively in relation to the art piece which was mentioned in our 10 post sample, and appears prominent in these discussions.

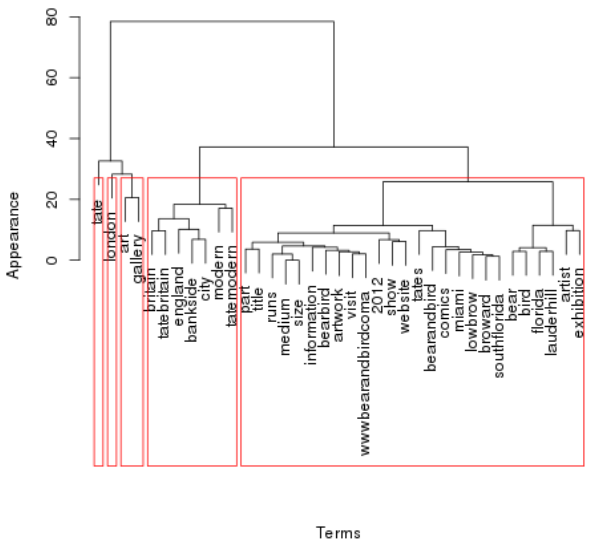
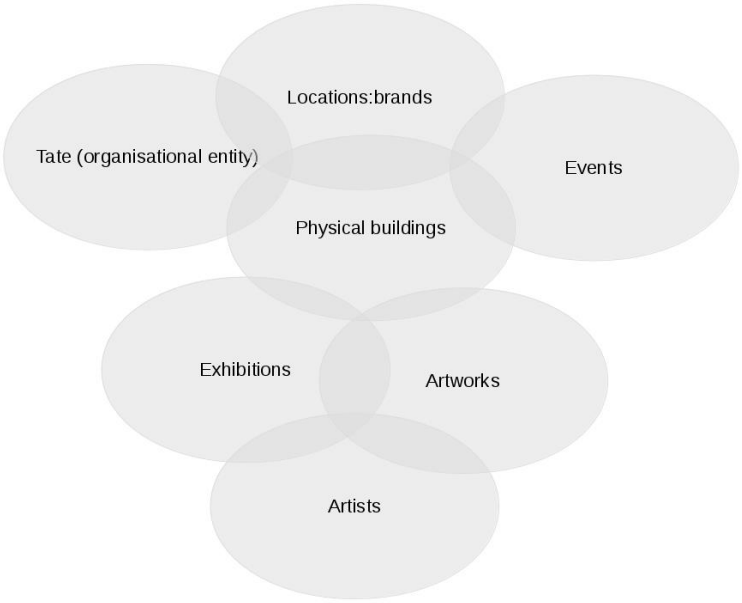


Fig. 4-28. Dendrogram of Flickr terms relating to Tate.

The Flickr dendrogram (see Fig. 4-28) shows terms relevant to various Tate institutions (Britain, Modern, London, [St.] Ives). However, we also find an additional sense 'Tate' identified, in particular a cluster relating to a popular comic shop (called 'Tate') in Florida which also hosts exhibitions. Although this sense of the term appears relatively infrequently, this large cluster is consistent in theme and subject and therefore appears prominently in the data.

A Social Media Entity-level View of Tate

Social media can be mapped to the Tate entity set in the following manner:



Social media comments may relate to each entity: critique of Tate as a brand, responses to practices or collections, attendance and response to specific events, which may be identified through social media trending topic or event detection methodologies (e.g. [Becker et al., 2011; Zubiaga et al., 2011]), and characterised through approaches such as sentiment analysis of social media (e.g. [Pang & Lee, 2008]). We also note reactions to Tate's physical placement, accessibility, etc.; responses to

exhibitions, artworks, and to the artists themselves. Additionally, evaluation of social media data using social network analysis techniques can permit characterisation of that community, and (usefully from the perspective of change management/appraisal of semantic change), of the change(s) in communities over time.

Discussion and Conclusion

During this initial evaluation, we collected data relating to aspects of Tate activity from three social media sources: Twitter, a microblogging site; Flickr, a photograph sharing website; Tumblr, a website that combines the features of microblogging and media sharing. Following an initial aggressive filtering step, we employed qualitative evaluation to explore the dataset through limited numbers of randomly chosen exemplars.

The quantitative exploration of the data involved an initial development of a vector space (matrix) representation of each dataset, extracting key terms and exploring their semantic neighbourhood enabling the identification of terms which correlate most closely. The quantitative survey demonstrates that the term 'Tate' is applied on entities unrelated to the Tate brand: in particular, the Twitter dataset contained a number of references to an adult film star of that name, whilst the Flickr dataset contained a number of references to the Tate comic shop in Florida, USA. Within a vector space model, however, it is relatively easy to disambiguate these separate referents.

Findings from this study show differing focus, coverage and informational content of material between the social networks reviewed. This suggests that a greater number of social network resources is likely to increase overall coverage of online communities.

To summarise, we would expect social media to play a role in two main ways: First, social media is a useful source of information relating to metrics associated with appraisal such as impact and reuse of material, perspectives on entities (such as artists, institutions and individual artworks) contained within or of relevance to a catalogue. Social media sources can usefully provide source material for evaluation of 'alt-metrics' (see for example [Priem et al., 2010]) in relation to visitor access and use of resources and art objects. As Thelwall et al. (2013) note, coverage is low (that is, the majority of individual catalogue items held within Tate are not specifically referenced, as indeed might be expected); however, coverage is present of a number of aspects of Tate's activities and resources that are not traditionally considered as holdings, such as perspectives on collections, collection policy, exhibitions, activities or buildings.

Secondly, we note an additional potential use in supplementing catalogue information, although given the main content of these social media resources, we presently view this as a more challenging usage, requiring further evaluation and development to fully realise.

Ontology Population

This subsection now focuses on populating the Art & Media domain ontologies with instances coming from external sources. Ontology population is part of the "umbrella" notion of *ontology learning*, which refers to the automatic or semi-automatic construction, enrichment and adaptation of ontologies [Maedche & Staab, 2004]; other core inherent processes include *ontology enrichment* (i.e. extending an existing ontology with additional concepts and semantic relations) and *inconsistency resolution* (i.e. resolving inconsistencies with the view to acquire a consistent (sub)ontology). It should be noted that the process of ontology population does not change the structure of an ontology but only the set of realisation (instances) of concepts and relations in the domain. Filling an ontology with accurate and comprehensive instances, has been proven useful in many applications, including dictionary construction, search query refinement, automatic question/answering, system-based decision support, etc.

In this context, we have developed *PROPheT* (*PeRicles Ontology Population Tool*), a novel GUI-equipped instance extraction engine, responsible for locating instances (realisations) of concepts and relations in a Linked Data source, filtering them and subsequently inserting them into an OWL-based domain ontology. To the best of our knowledge, no tool exists that can offer the extent of functionality delivered by PROPheT (see "Comparison with other Tools" section later on).

For the purposes of this deliverable, all testing is performed on the Software-based Artwork (SBA) ontology, while DBpedia will serve as the Linked Data set. Nevertheless, the developed tool is flexible enough to seamlessly work with any domain ontology (written in OWL) and any RDF Linked Data set that is available via a SPARQL endpoint. The PROPheT tool, along with its source code and documentation, will be soon publicly released.

Background

DBpedia[18] is a crowd-sourced community effort to extract structured information from Wikipedia and to make this information available on the Web [Bizer et al., 2009]. The structure and the content of this central interlinking hub for the emerging Web of Data makes it easier for the huge amount of information originating from Wikipedia to be used in some new interesting ways.

The term "Linked Data" refers to structured and interlinked data shared in a way that can be read automatically by computers. This enables data from different sources to be connected and semantically queried. The Linked Data paradigm builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, it extends them to share information.

The UI development of PROPheT was based on Python[19] along with Qt[20], and specialised Python APIs and libraries (rdflib[21], SPARQLWrapper[22]) were integrated to enable handling of local and remote models and content. Additionally, an SQLite[23] database was set to store dynamic data that are created/manipulated through the operation of PROPheT.

PROPheT Components and Features

The PROPheT tool populates a target (domain) ontology with instances retrieved from a Linked Data source. The general components defined in the context of PROPheT are the following (see also Fig. 4-29):

- **My Model** (MM) is the ontology to be populated with new instances. The ontology file must comply with a specific format (.owl, .rdf, .ttl) and can reside at a local or remote location.
- **External Model** (EM) is the source from which new instances will be retrieved. This source should be facilitated by a SPARQL serving mechanism (i.e. endpoint URL), so that PROPheT will be able to run SPARQL queries directly to the source.
- **PROPheT extraction module (search mechanism)**: The current version of PROPheT features class-based and instance-based population; for more details see next subsection.
- **PROPheT mapping module**: allows the user to match MM and EM datatype properties. It stores and utilizes these user-defined matchings to instantiate existing datatype properties of populated instances with values derived from EM.

[18] http://wiki.dbpedia.org/
[19] https://www.python.org/
[20] http://www.qt.io/
[21] https://github.com/RDFLib/rdflib
[22] https://github.com/RDFLib/sparqlwrapper
[23] https://www.sqlite.org/

- **PROPheT storage module (database)** stores PROPheT preferences relevant to setting limits to the number of derived results from EM, adding/removing MM and EM sources, adding/viewing/deleting user-defined mappings, etc.
- **PROPheT export module**: A mechanism for storing the already processed/populated MM in a local file. Allowed formats are: .owl, .rdf, .ttl, .nt, .nt3.

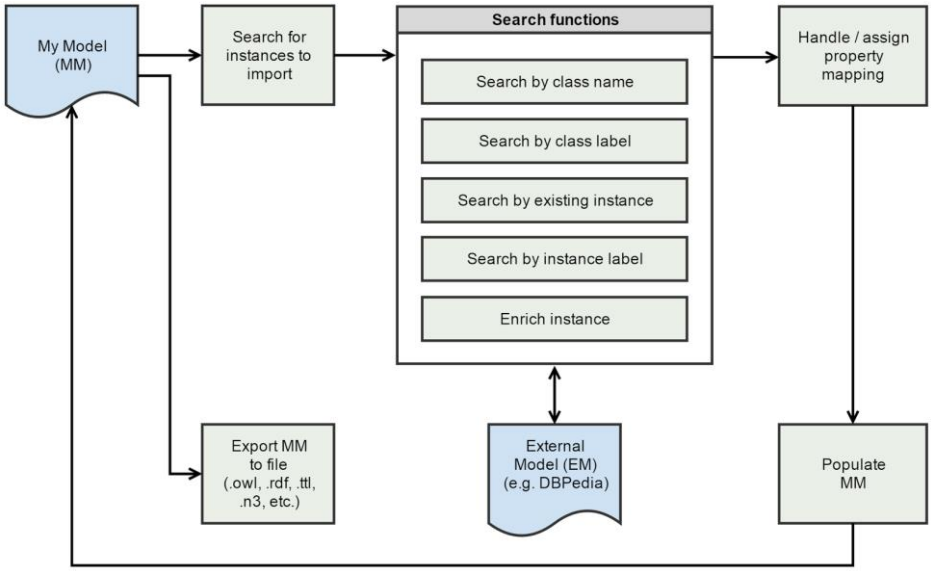


Fig. 4-29. PROPheT main workflow.

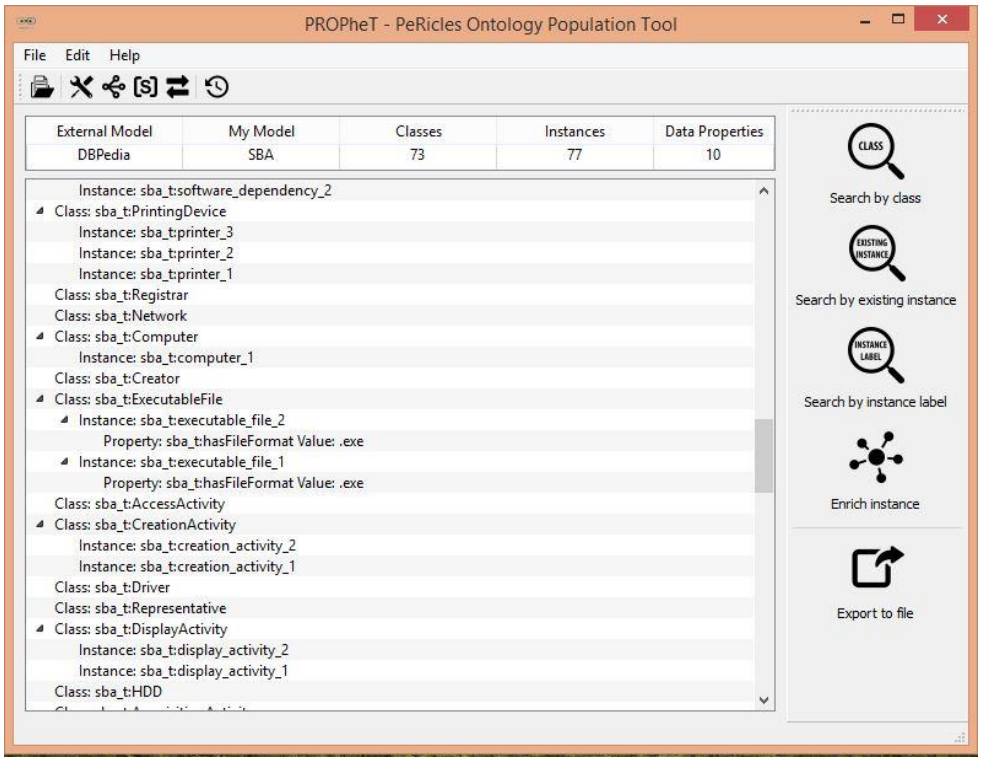


Fig. 4-30. PROPheT main window.

PROPheT's main window is presented in Fig. 4-30, where *DBPedia* is defined as the EM and *SBA ontology* is loaded as MM to the software. As can be seen, there are 73 classes, 10 data properties

and 77 instances defined in the specific version of the SBA ontology loaded in PROPheT. The software visualises these three types of information in a tree diagram, where *parent nodes* are classes defined with a "prefix:classname" form in the tree, and *child nodes* are instances that are included in the referenced class, together with their datatype properties and values, placed as leaves in the tree.

PROPheT Functionality

PROPheT offers three types of instance extraction-related functionalities:

1. **Class-based populating** enables the user to populate new instances "from-scratch" in MM, by searching for specific types of classes in EM. The class-based search may be performed in two different ways:

   a. *Search by class name* - by entering, if known, the exact type of class, for example `dbo:Artist`[24], or
   b. *Search by class label* - by entering a text that corresponds to the label (`rdfs:label`) of a class, e.g. "artwork", that we want to search in EM. This approach is more flexible for cases when the user does not know the exact structure of EM or when the class type belongs in the hierarchy of the ontology.

   In both cases, the software performs SPARQL queries to the EM endpoint and tries to detect matches between the user-defined class and classes that exist in EM. As a result, a list of corresponding instances that belong to the specified class is returned, if applicable. The number of responses from EM may be limited by setting the value of the corresponding parameter in the preferences window. The user may then select the instance(s) that he/she wishes to import (populate) under an existing MM class.

   In order for PROPheT to proceed with the defined instantiation(s) for MM, ontology mapping should be performed. For that reason, a list of all unique datatype properties (`owl:DatatypeProperty`) for the selected instance(s) is given to the user in order to consistently define their mapping into existing datatype properties in MM. The user-defined mappings are stored/updated in a local database in PROPheT and, additionally, are taken into account for proper instantiations of the new individuals (instances) in MM.

2. **Instance-based populating**: Detecting and importing new instances with an instance-based search may be done in two different ways. These are:

   a. *Search by existing instance* - The user may select an instance already existing in MM and query the endpoint for similar instances. In detail, PROPheT performs an `rdfs:label`-based search and finds EM classes that include an instance with the desired instance's label. The user may then select the classes of his/her interest, view their collection of instances and choose which of them will be imported to MM.
   b. *Search by instance label* - Similarly to the previous technique, a straight `rdfs:label`-based search is executed. However, in this case the user needs to type the desired label and the search will result in a set of instances associated with this specific label, rather than similar instances of the same type (class).

   In both instance-based mechanisms, the software presents a list of search results (instances) which may be selected by the user and populated in MM. A mapping process needs to be completed, as described in class-based populating section above, in order for the tool to also import the datatype properties of instances and their corresponding values.

3. **Enrich existing instance**: This type of functionality gives the user the ability to enrich an already existing instance within MM with properties and values derived from other instances in EM that

[24] dbo is the prefix for a specific DBPedia URI, that is http://dbpedia.org/ontology

have the same `rdfs:label`. By selecting an existing instance, PROPheT checks if a label for this instance is defined and performs an `rdfs:label`-based search to find instances in EM that include the desired label. Potentially, the derived instances may belong to one or more different classes in EM; the software tracks and presents to the user the defined `rdf:type` property declarations for these instances. Based on the content and the semantics of the derived instances, the user may decide which pair(s) of property-value he/she wants to import to the MM for the initially selected instance. Similarly to class-based and instance-based populating functions, an ontology mapping process should be performed in order for the new properties and values to be added in the existing instance.

Comparison with other Tools

For the process of ontology population (and, more generically, ontology learning), several approaches and practical tools have been proposed in literature, differing in the following categorisation criteria [Gómez-Pérez & Manzano-Macho, 2003; Petatsis et al., 2011]:

- **The knowledge base**, which is the source of information and can be any text, web content, thesaurus, dictionary, database, ontology, linked data (LOD), etc.
- **Degree of automation**, meaning the degree that a system/tool automates decisions or else the user intervention in the process. Various degrees of automation may be identified in different modules of the system.
- **Featured tasks and functionality**, meaning methods related to learning or knowledge acquisition, information extraction, merging content and the final outcome.
- **Initial requirements**, such as prior knowledge, type of required input for learning and populating an ontology, availability and portability of resources, etc.
- **Efficiency of results**, dealing with the evaluation of the ontology population process, and how the new set of realisation (instances) of concepts and relations can be the optimum one.

We have selected some representative ontology population tools from the available ones in literature, to perform a qualitative comparison, which is summarised in Table 4-5.

Table 4-5. Analysis of existing ontology population tools.

Tool	**DB2OWL [Ghawi and Cullot, 2007]**
Resource/input type (S/SS/U[25])	relational database (S)
Output	ontology (OWL)
Functionality	• SPARQL query driven population • instance export
Degree of automation	automated mapping definition and instance export
User Interface (GUI)	yes
Required user knowledge	• domain • mapping • evaluation
Additional features	uses semantic relation for database-to-ontology mapping

[25] S: structured, SS: semi-structured, U: unstructured

Tool	[Gavankar et al., 2012]
Resource/input type (S/SS/U)	DBPedia or other LOD sources (S)
Output	ontology (no specified type)
Functionality	• SPARQL query driven population • mapping between subset of predicates of an academic ontology and those in DBPedia or other LOD sources
Degree of automation	manual mapping using external data sources
User Interface (GUI)	information not available in paper
Required user knowledge	• domain • mapping • ontologies
Additional features	• domain oriented (academic ontology) • handles object and data properties

Tool	SPRAT [Maynard et al., 2009]
Resource/input type (S/SS/U)	text documents (U)
Output	ontology (no specified type)
Functionality	uses patterns for NE (named entity) recognition and aligns them into concepts of instances
Degree of automation	automated (linguistic pre-processing of text, NE recognition via machine learning and rule-based methods)
User Interface (GUI)	yes (GATE used)
Required user knowledge	• ontologies • text mining
Additional features	• combines aspects from traditional NE recognition, ontology-based information extraction and relation extraction • may create a new ontology from scratch

Tool	[Jeong, 2013]
Resource/input type (S/SS/U)	relational database (S) or web text (U)
Output	ontology (no specified type)
Functionality	• proposed web-based learning and population model • mapping is based on system's recommendations via rule-based reasoning

Degree of automation	semi-automated: candidate instances via rule-based learning operation
User Interface (GUI)	yes
Required user knowledge	• domain • ontologies • mapping evaluation
Additional features	• performs ontology generation and population • model based on an ISO related to a Metadata Registry MDR (standardise semantics) • handles object and data properties

Tool	**RDF123 [Han et al., 2008]**
Resource/input type (S/SS/U)	spreadsheet (Google sheet or CSV file) (SS)
Output	ontology (RDF)
Functionality	mapping via user-defined alignment
Degree of automation	semi-automated: user constructs a graphical rdf template that specifies how each spreadsheet data or metadata is converted to an rdf node
User Interface (GUI)	yes (application and web service)
Required user knowledge	• domain • ontologies • mapping
Additional features	rdf template is stored as a valid rdf document encouraging reuse and extensibility

Tool	**Mid-Ontology [Zhao and Ichise, 2012]**
Resource/input type (S/SS/U)	LOD sources (S)
Output	ontology (no specified type)
Functionality	• SPARQL query driven research and population • takes into account the owl:sameAs property • based on predicate grouping
Degree of automation	automated retrieval of related information from LOD and integration with target schema
User Interface (GUI)	information not available in paper
Required user knowledge	• domain • ontologies

Additional features	• constructs an (intermediate) ontology from scratch • integrates predicates from different data sets -performs term extraction and similarity matching • removes noisy instances (those with non-valid data: broken links, empty fields, etc.)

Tool	**OntoLearn [Velardi et al., 2002]**
Resource/input type (S/SS/U)	text from specialised web sites (U)
Output	information not available in paper
Functionality	• extracts terminology from a corpus of domain text • filters terms via NLP (natural language processing) and statistical techniques • uses WordNet to perform semantic interpretation
Degree of automation	automated extraction and integration of concepts in domain ontology
User Interface (GUI)	information not available in paper
Required user knowledge	evaluation
Additional features	enriching a domain ontology with concepts and relations

Tool	**OntoBuilder [Modica et al., 2001]**
Resource/input type (S/SS/U)	data encoded in XML or HTML (SS)
Output	ontology (in XML)
Functionality	has two phases: 1. training (the initial ontology is built using data provided by user), and 2. adaptation (refinement and generalization of initial ontology (with additional concepts and values)
Degree of automation	semi-automated processes: the user suggests additional sources of information and evaluates every candidate ontology extracted and merged with existing one
User Interface (GUI)	yes
Required user knowledge	• ontologies • evaluation
Additional features	adopts text mining and NLP techniques to perform ontology adaptation and enrichment

Tool	**SAPOP [Sun et al., 2013]**
Resource/input type (S/SS/U)	LOD sources (S)

Output	ontology (no specified type)
Functionality	proposes candidate instances per category via a semi-automatic/semi-supervised learning method
Degree of automation	semi-automatic framework for discovering reliable seed instances and evaluating results
User Interface (GUI)	no (framework presented)
Required user knowledge	evaluation
Additional features	• refinement cycles help the efficiency of the seed discovery and labeling process • identifies the reliability of predicted instances

Most of the tools suggested in literature build an ontology from scratch while our PROPheT approach aims to enrich an existing ontology with relevant instances and values of represented concepts, without any restriction to its portability to new thematic domains. To the best of our knowledge, there is no other ontology population tool that instantiates new concepts (individuals) from a Linked Data (LOD) source to an ontology, regardless of the domain of interest or the content of LOD. The flexibility of our proposed tool lies in the fact that any kind of LOD with a served endpoint can be handled by PROPheT as an external source of knowledge for extracting concepts of interest and populating them to corresponding resources in the domain ontology needed.

In our approach, we take advantage of the nature of data included in the knowledge base; unlike pure text documents, information in LOD is stored and can be manipulated in a structured, semantically enriched way. Since we deal with ontologies, we conserve the structured nature of data for both the input (the one to be populated) and the output (the derived one) ontology, which can comply with any of the main ontology file types (.owl, .rdf, .ttl, .nt and .n3.).

Concerning the degree of automation, according to the conducted literature review, there is no implemented tool that carries out the whole population process automatically. PROPheT may be considered as a semi-automatic, user-driven system with different degrees of automation in various tasks: the interaction of the software with the LOD endpoint for performing the search process is done automatically, while the final selection of the instances to be populated in the ontology and the mapping of properties needs to be done manually. The user-driven mapping process enables the dynamic and error-free definition of matching elements between source and target ontology.

The intervention of the user/expert is critical for the use of PROPheT in order to avoid containing contradicting information which may lead to an inconsistent ontology. The user needs to know the domain that the initial ontology describes, in order to perform proper selections in the mapping process. However, the required user expertise does not include any knowledge regarding the way of communicating/querying the endpoint, or of the process of populating instances to the ontology in practice.

Finally, in the current version of the software, elimination of redundancy in the instance set is handled adequately (duplicated instances cannot exist in the ontology), while semantically similar instances could be automatically identified and removed in future versions.

Limitations and Future Work

The current version of PROPheT does not handle direct or indirect imports of modular ontologies. The user-defined inference level of imported ontologies is intended to be a feature of PROPheT.

Based on the inference level, the software will handle the domain ontology of MM as being an extended version of an ontology graph, where declarations of all triples up to the inference level of imported ontologies are combined with those triples that are defined only in the domain ontology.

An interesting aspect of ontology population, which is not addressed adequately in the literature, is the handling of *redundancy*, referring to the process of identifying if instances of the ontology refer to the same real object. The latter entails risks, since the redundant instances may contain contradicting information, which may in turn lead to an inconsistent ontology. Redundancy in the instance set can be resolved by *entity disambiguation.* This process may enhance the plausibility of an ontology by facilitating the process of querying the ontology, and at the same time by limiting the size (and complexity) of the ontology [Petasis et al., 2011]. PROPheT merely handles instance redundancy in a way that instances with the same name-identifier cannot be populated multiple times in the ontology; values of populated data properties are linked to one instance. On the other hand, PROPheT does not currently encompass more complex handling mechanisms, such as heuristics or machine learning methods that identify similar resources, but this task is appointed to be fulfilled as future work.

The degree at which a system automates processes and decisions is another important aspect of population ontology tools. PROPheT may be considered as a semi-automatic user-driven system with different degrees of automation in various tasks. The interaction of the software with the EM endpoint for performing the aforementioned search functionalities is done automatically, within the context defined by these functionalities. However, the final selection of the instances to be populated in the ontology and the mapping of EM and MM datatype properties need to be done manually by the user. The required user expertise does not include any knowledge regarding the way of communicating/querying the endpoint, or of populating instances to the ontology in practice; nevertheless, the user needs to know the domain that the MM ontology describes, in order to do proper selection in the mapping process.

The type of populated information is also significant for the completeness of an ontology population tool. Instances in ontologies may contain values for both object properties (`owl:ObjectProperty`), which relate individuals (instances) of OWL classes with other instances, and datatype properties (`owl:DatatypeProperty`), which relate individuals of OWL classes to literal values. In the case of PROPheT, it is not possible to populate object properties; relationships between instances are a more complex matter since each relationship is limited by domain and range attributes and the manual alignment between MM and EM properties would mostly lead to inconsistent results.

4.3. Chapter Summary

The chapter presented our novel proposed prototype methods for extracting semantic information from visual and text-based content. More specifically, **SALIC was presented, a prototype method for detecting semantic concepts from the content of a DO** that (a) automatically gathers training data with minimal annotation effort, (b) minimises the number of required training instances, and, (c) increases the performance of the classification models by utilising a smart sampling approach. **PCS was also presented, which is a novel and very fast method for dimensionality reduction**.

Further we introduced a tool for the **scalable, supercomputer-friendly processing of semantic media content based on vector fields** which paves the way for the exploration of quantum-like content behaviour in digital collections, including the typology of correlations underlying machine learning for automatic content analysis.

With regards to semantic information extraction from text-based content, the deliverable also presented our approaches for **analysing source text documents** relevant to the two case studies and using the extracted information for **populating the developed domain ontologies with instances**.

5. Extraction and Analysis of Use Context Information

The most widely referenced definition of context is given by Dey et al. (2001), according to which context is "*any information that can be used to characterize the situation of an entity. An entity is a user, a place, or a physical or computational object that is considered relevant to the interaction between a user and an application, including the user and application themselves*". We have looked at the different types of contextual information that can be gathered from the environment of use of Digital Objects in Deliverable 4.1 [PERICLES D4.1, 2014], the context of definition of Significant Environment Information (SEI) and the PET framework for information extraction. Our focus now shifts to a specific type of context, referred to as "*use context*", which refers to information related to contexts of use of the DO. This chapter discusses the adopted approaches for extracting and analysing use context information, in order to address issues like e.g. variations of DO interpretation, and to derive meaningful correlation links among content objects and use contexts.

5.1. Representation of Use Context

We start by presenting our adopted representation of use context. The modelling approach of PERICLES relies heavily on ontology-based constructs. Nevertheless, to the best of our knowledge, no approaches exist currently for ontology-based representation of *use context* in digital ecosystems. One can find several approaches in literature discussing formal context modelling via ontologies as a key means for representing context awareness in pervasive computing and distributed environments (e.g. [Chen et al., 2003; Ejigu et al., 2007; Schilit & Theimer, 1994; Wang et al., 2004]). The context in such environments may for example be defined in terms of location ([Weiser, 1993]), human factors (e.g., user habits, emotions, social environment, goals), or physical environment (e.g., location, infrastructure, noise, light, humidity) (see [Schmidt et al., 1999] for further description of context awareness). Although it may be challenging to incorporate such information into digital ecosystems, we note the potential of contextual information relating to the physical environment in particular as being relevant to curators in art and museum domains.

Alongside ubiquitous computing, the notion of context modelling has received attention in information retrieval. A search engine is provided with very few words yet is expected to retrieve relevant matches. To improve accuracy it is commonplace to make use of a user model containing a number of aspects of what in ubiquitous computing terminology would be viewed as human and physical environmental context. For example, what is the user's preferred language? Which previous queries have they made? What are their interests? What is their approximate location, and how may this influence their preferences? In particular, collaborative filtering approaches form particularly high-profile examples in which user profiling information is applied to match the user to a cluster of like-minded users, and hence to elicit statistical information from which to predict their 'likes' and 'dislikes' ([Moshfeghi, 2012]).

Context-awareness in information retrieval has received a great deal of academic interest over a number of years, with these relating to the relevance of context to information retrieval ([Cai et al., 2007]), to query interpretation and judgement of accuracy, adaption to query context ([Bai et al., 2007]), user belief ([Lau et al., 2008]) and social network cliques ([Gan et al., 2011]). Classically, interest in context-awareness in information retrieval often considers the question of how context and situational factors modify information-seeking behaviour ([Cool & Spink, 2002]). Context in information retrieval is likely to supplement considerations of physical location with 'cognitive, social and other factors related to a person's task, goals and intentions, which precipitate information

seeking episodes' ([Cool & Spink, 2002, pp.605-606]). What's more, the knowledge available to a user is also likely to be a factor, although it is seldom explicitly modelled. Information on user expertise and community engagement may be used to classify potential query matches ([Deng et al., 2009b]). In general social context is increasingly viewed as relevant to information retrieval, especially those in which subjective judgements of relevance are key.

In addition, context has been studied in other environments as well, including e-Learning (e.g. [Dichev & Dicheva, 2005; Derntl & Hummel, 2005]), where the notion of context is explored for improving the performance of content search and retrieval, and Web Services (e.g. [Maamar et al., 2006; Mrissa et al., 2009]), where context helps characterise the interactions between humans, applications, and the environment. Interestingly, specifically the study of use context has been sporadically investigated, with the work by Jovanović et al. (2007) being such an example. The authors attempt to formalise the specific context of use of Learning Objects (LOs), in order to optimally personalise the learning process by tracking the usage of LOs. We believe this to be very significant, since in in many recent theories about learning, notably those influenced by constructivism ([Vygotsky, 1980; Piaget, 1970]) such as discovery learning ([Bruner, 1961]), the student is viewed as making use of existing knowledge to contextualise new information, to in turn construct new knowledge. We note that a nuanced view on such educational theories is advisable (cf. [Phillips, 1997; Mayer, 2004]), nonetheless there is broad acceptance of the significance of individual learner profiles in supporting the learning process. As such, we would expect teaching to be learner centred, with new learning material relating to the relevant pre-requisites already held by the student, whether this is learning objects or descriptive texts in self-directed learning. In the case of this latter example, we note the ILEX system which modelled existing knowledge of the learner in order to present new information (labels of museum objects) in a relevant and targeted way in order to improve their learning experience ([O'Donnell et al., 2001]). Involved notions include LO domain topic, activities in which a LO is used, references to learners and so on. Several of these notions - but with required justifications - are deployed also in our domain ontologies. Another similar example is the work by Strang et al. (2003), where the authors are using (a) context for tracing the execution of Web services, and (b) past context to predict and adapt the behaviour of Web services. All in all, a more thorough survey of context and use context modelling approaches will be given in the upcoming deliverable D4.4 due M40.

5.1.1. Key LRM Resources

The key resource types that were adopted from the LRM [PERICLES D3.3, 2015] by the domain ontologies are:

- **Abstract resource** (`lrm:AbstractResource`) - this class models resources that have no physical extension but may have a temporal extension or may be subjected to change. This class was further specialised in the domain ontologies for representing e.g. conceptualisations of artworks, descriptions, activities, etc.
- **Concrete resource** (`lrm:ConcreteResource`) - this class models resources that have a physical and temporal extension. As such, they can be located in space and time, and the nature of their physical extension may be described accordingly. Again, this class was further specialised in the domain ontologies for representing digital videos, equipment (computers, storage devices), etc.
- **Aggregated resource** (`lrm:AggregatedResource`) - this class is aimed at modelling parthood relations (via the property `lrm:hasPart`), namely, relations of part-to-whole and relations of part-to-part within a whole.
- **Agent** (`lrm:Agent`) with subclasses the human agent (`lrm:HumanAgent`) and software agent (`lrm:SoftwareAgent`) - the class models the main resource that, based on its actions, performs activities and triggers events; an agent is a bearer of change in the ecosystem. Human

agents are further specialised in the domain ontologies, like e.g. artists, creators, programmers, staff, etc., and software agents into programs, software libraries, operating systems, etc.

- **Dependency** (`lrm:Dependency`) - this class indicates the requirements for some conditions (or resources), in order for a resource (`lrm:Resource`) to fulfil its purpose or to function adequately and consistently. Dependencies usually connect one or several resources (designated by the `lrm:from` property) to one or several other resources (designated by the `lrm:to` property). Additional properties attached to a dependency are intention, specification, impact and precondition; Fig. 5-1 displays the representation of a dependency.

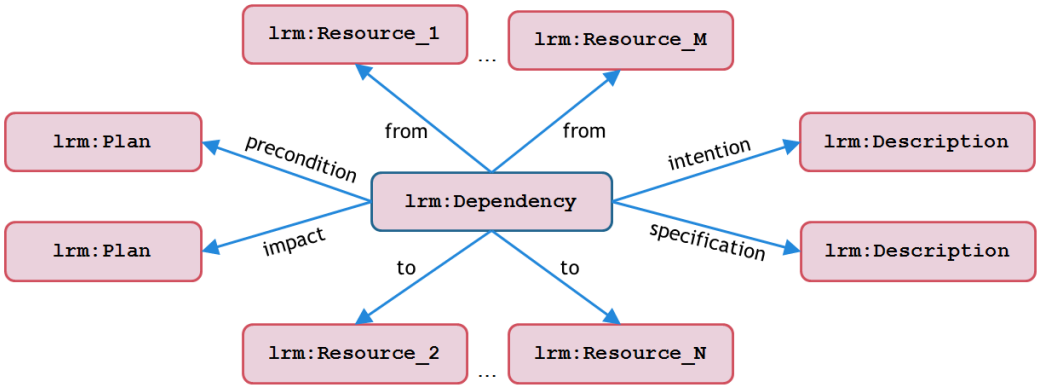


Fig. 5-1. Dependency representation with related properties.

For the domain ontologies and in order to model complex relationships between resources within the context that the domain defines, we extend the basic notion of `lrm:Dependency` with three sub-categories of dependencies, the details of which are given below:

- **Hardware dependencies**: specify hardware requirements for a Resource, or relations among them, in order for it to function properly.
- **Software dependencies**: indicate the dependency of a Resource or Activity on a specific software (Software Agent) - name, version, etc.
- **Data dependencies**: imply the requirement of some knowledge, or data or information, in order for a Resource to achieve its purpose of existence or function. This kind of data may originate from human input (e.g. file passwords), computer files (e.g. configuration files), network connection, live video, etc.

5.1.2. Representing Context and Use Context

Context in the domain ontologies is expressed via the associations between key classes `lrm:Agent`, `lrm:Activity` and `lrm:Resource`, which are shown in Fig. 5-2. In detail, agents are related to activities via property `lrm:executes` and its inverse property `lrm:executedBy`. Furthermore, when relating an activity with a resource, the latter can be either (a) the resource that is affected by the activity (or that the activity targets at), or (b) a resource that was used during the activity execution. In other words, a target resource is the one mainly handled by the activity (e.g. created, borrowed, destroyed), while used resources are those manipulated for the activity execution (e.g. equipment, software, hardware, etc.). The first type of relationship is indicated by object property `:targetsResource` (inverse of `:targetedByActivity`) and its specialised subproperties, whereas the second type is represented via property `lrm:used` (inverse of `lrm:usedBy`). Last, as seen in Fig. 5-2, there is no direct connection between agents and resources, since they are only associated via activities.

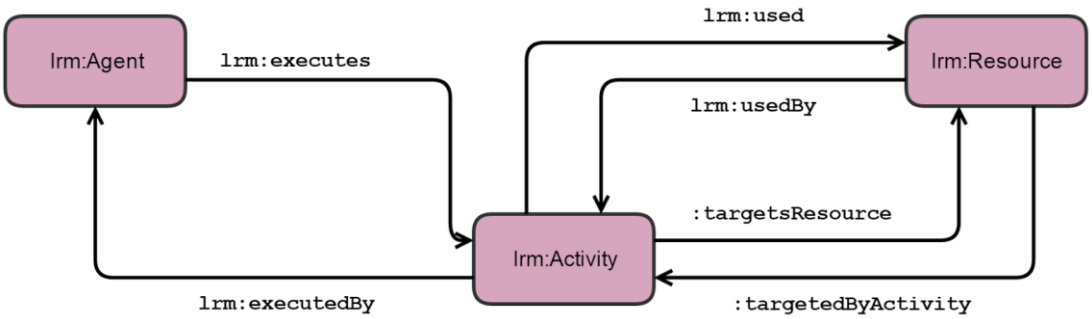


Fig. 5-2. Associations between key classes in A&M domain ontologies.

Despite the availability of the aforementioned properties, relationships may also be modelled with dependencies, as described in the previous section, should it be necessary to further describe the relationship itself with properties such as intention, specification, etc.

Regarding the representation of the use context, as mentioned above, the use context of dependencies between resources is additionally enriched in LRM with the notions of **intention** that specifies what a dependency intends to express, and **specification**, which thoroughly describes the dependency itself and its context. Additionally, the LRM defines two more dependency descriptors (object properties), named **precondition**, which describes the contextual properties that need to hold in order to consider the dependency as "activated", and **impact**, which describes what must be done when the dependency is activated [PERICLES D3.3, 2015]. In other words, dependencies with defined intentions, specifications, preconditions and impacts constitute **meaningful correlation links among resources and use contexts**.

Specifically for the Art and Media (A&M) ontologies, the extension of `lrm:Dependency` is enriched by adding a set of predefined intention types for seamlessly representing all relevant dependency occasions; these are **conceptual**, **functional** and **compatibility** intentions, the detailed description of which is given below:

- **Dependencies with a conceptual intention** are aimed at modelling the intended "meaning" of the resource (i.e. artwork) by its creator, according to the way he/she meant for the artwork to be interpreted/understood. For example, a poem (digital item) belonging in an archival record may not conserve its formatting during the normalisation process, something that is against the intention of the artist regarding the way that the poem is conceptualised/conceived by a reader.
- **Dependencies with a functional intention** represent relations relevant to the proper, consistent and complete operation of the resource. For example, a specific type of monitor is required for properly displaying a software-based artwork.
- **Dependencies with a compatibility intention** model compatible software or hardware components which may operate together or as substitutional components for availability, obsolescence or other reasons. For example, the software used for playing back a digital video artwork consistently, is compatible and may be substituted with other software items that demonstrate the same functionality.

5.1.3. Instantiations of Use Context Representation

This subsection provides instantiation examples to demonstrate how the representation of use context in the A&M domain is achieved through dependency descriptors such as intention and specification. The first three examples (Fig. 5-3 to Fig. 5-5) implement a real-case scenario of a specific software-based artwork named *Becoming*[26] (created by Michael Craig-Martin, in 2003), as

[26] http://www.tate.org.uk/art/artworks/craig-martin-becoming-t11812

described in [Falcão, 2010], while the rest of the implementations are derived from similar cases in the other A&M subdomains.

First, Fig. 5-3 displays a data dependency of the compilation activity of *Becoming* on the source code and on a set of images. According to the specification, in order for the compilation activity to succeed, the existence of these resources (source code and images) is necessary. Furthermore, the intention is of **functional** type, which means that the activity would not take place at all, if the dependency was not satisfied.

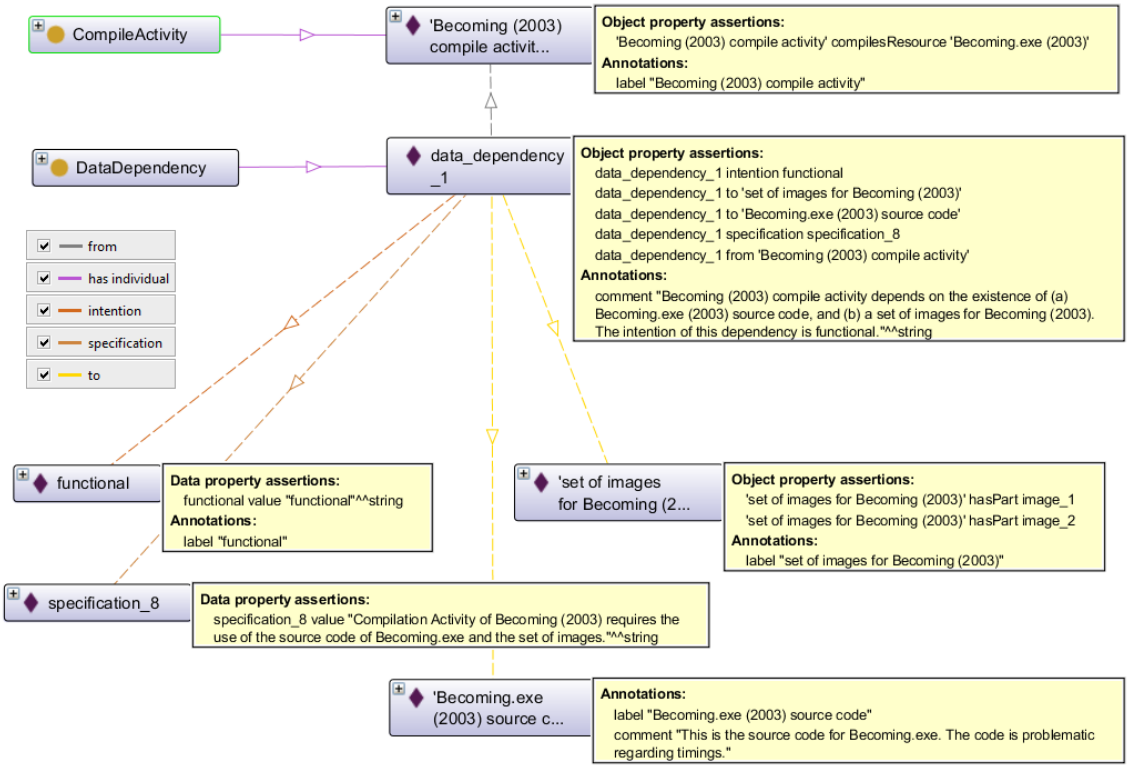


Fig. 5-3. Instantiation of a functional intention.

Fig. 5-4 demonstrates a hardware dependency with **conceptual** intention. Referring again to *Becoming*, the dependency specification determines that the resource (artwork's hardware) should look "clean and slick", a prerequisite set by the artwork's creator. The conceptual type of intention indicates that, even if the dependency was unsatisfied, all resources would still be usable and functioning but the overall result would not satisfy the artist's initial intention. Therefore, the conceptual interpretation of the artwork would be inconsistently preserved, signalling a failure in the digital preservation process.

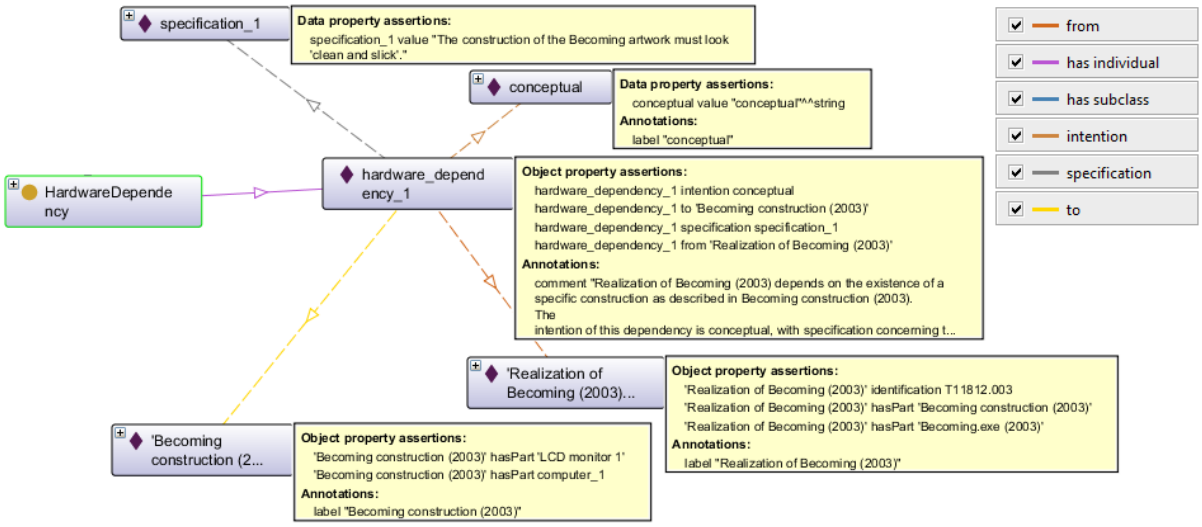


Fig. 5-4. Instantiation of a conceptual intention.

In [Falcão, 2010] it is described that the executable file of the *Becoming* artwork is dependent on Windows XP (Home Edition) x86. In other words, the display activity of the specific artwork through the use of the executable file is dependent on the aforementioned version of operating system (OS). Given the fact that any programme developed in an x86 OS can also be executed in an x64 OS, this software compatibility can be expressed via the software dependency with compatibility intention as seen in Fig. 5-5. A compatibility intention along with a software dependency specifies whether a resource can function or be used alongside with another resource or set of resources. Such a representation enables version integrity checking and change management validation, setting a use context for a specific resource.

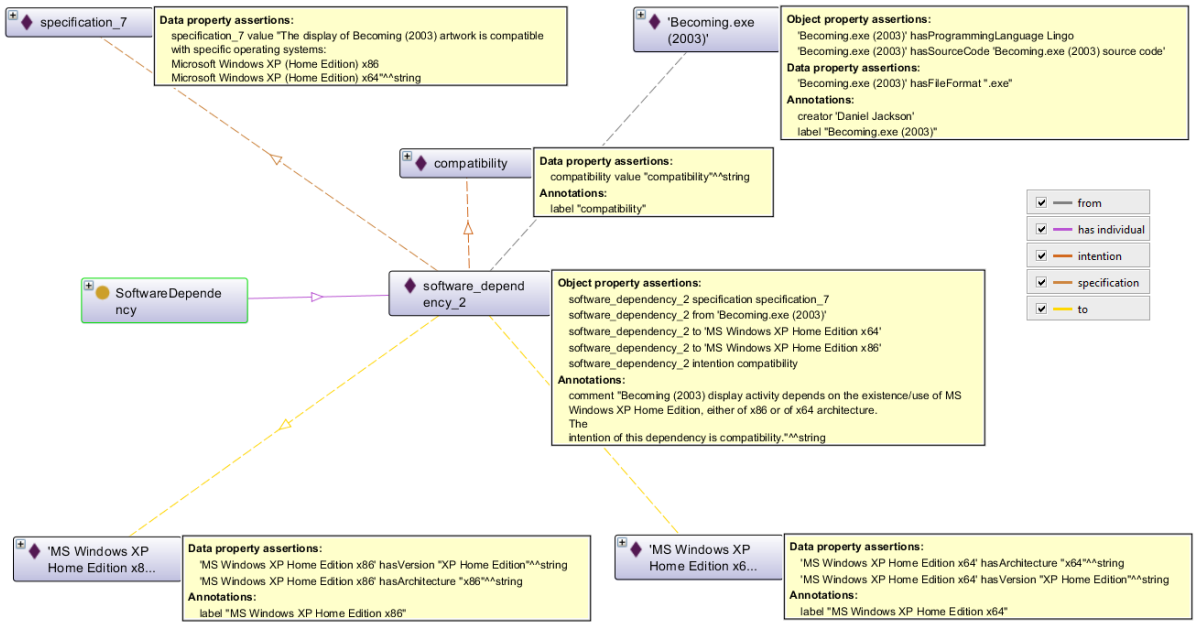


Fig. 5-5. Instantiation of compatibility intention.

A more complicated scenario is derived from the normalisation activity that may be applied in a digital item, for example a text file, in the born-digital archives (BDA) subdomain. Through the normalisation process, an access file is created from the initial one; the original file is in the format used by the creator, while the access format of the created file is defined in the archival policy

followed by the normalisation software used (e.g. OpenOffice). In terms of the BDA ontology, this instantiates a software dependency of the normalisation activity to the used software, while, apart from that, there is also a hardware dependency of the normalisation software to the hardware required in order for the software to run. The overall normalisation process is dependent also on the existence of the initial text file, and this information can be presented through a data dependency from the activity to the file, as seen in Fig. 5-6. The intention of all three types of dependencies appearing in Fig. 5-6 is functional, meaning that all the required resources modelled in this example impact the functionality of the resources for whom the dependencies were implemented (see `lrm:from` property).

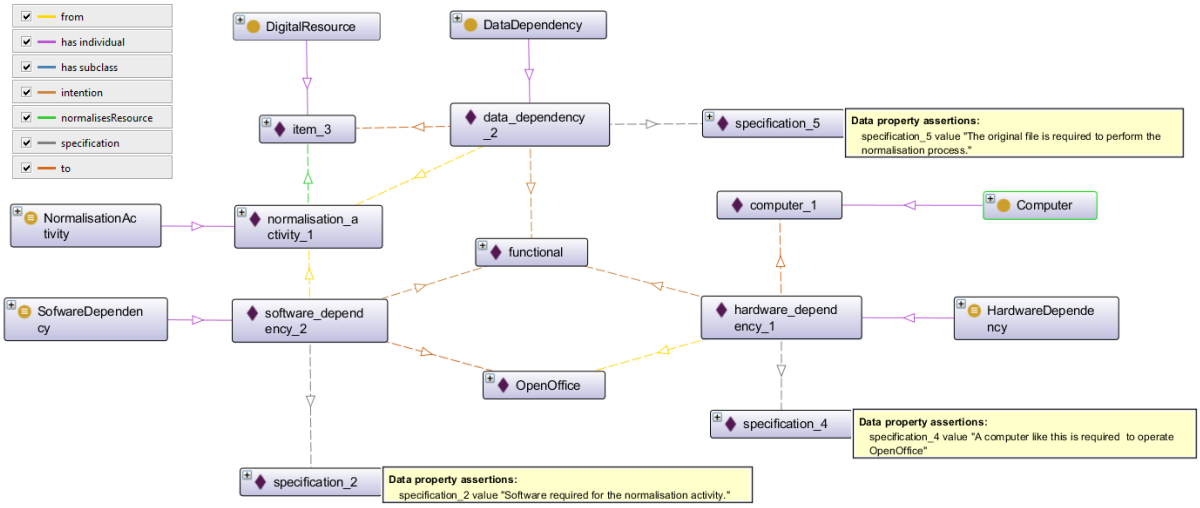


Fig. 5-6. All three types of dependencies combined together to attribute a complicated scenario of dependencies with functional intention.

For the sake of brevity, in the aforementioned figure, all hardware requirements are summarised as a computer system (*computer_1*) capable of running the software. However, the example could be easily extended to show hardware dependencies from specific components (such as CPU speed, RAM space, etc).

5.2. Analysis of Use Context Information

The PET (Pericles Extraction Tool) has been developed as an output of task T4.1, during the first months of the project (M18). Its aim is to extract useful information from the environment of execution and use of Digital Objects, along with general environment information. As PET was one among the first PERICLES outcomes, the output of the tool was kept in a general, rather free form (not adhering to a specific schema) format based on JSON. This was due to the fact that the ontologies for PERICLES were still in early development; and that PET output required further processing.

In order to embed the PET data into digital objects, a **Version1** of the PET2LRM service was developed in the scope of task T4.2; the component performs a transformation from PET data to a more structured, LRM-based form, that can be used to support embedding into DOs (as in task T4.2) and further analysis (this and future tasks T4.4 and T4.5). PET2LRM performs a direct mapping of PET information into simple LRM constructs. The implementation developed in T4.2 was capable of converting PET extracted environment information to LRM, starting from information stored in the original PET format and data files; and only for the information system (ecosystem) and file specific part of the PET data; that is excluding the event data captured by monitoring modules.

Since the development of PET2LRM, a more complete, PERICLES-specific ontology focused on the Digital Ecosystem concept is being developed (and is still under active development): the Digital Ecosystem Model (DEM) ontology. This ontology, together with a Java library (the EcoBuilder) to ease the creation of compliant models, has been reported in D5.2 [PERICLES D5.2, 2015], and will be further developed and reported in its final version in D3.5. Since the ontology has been created for PERICLES use, and presents constructs and entities that can be reported by the PET tool, we are planning to implement a new version of the PET2LRM tool, this time mapping to more complete and specific DEM entities, in the next WP4 tasks.

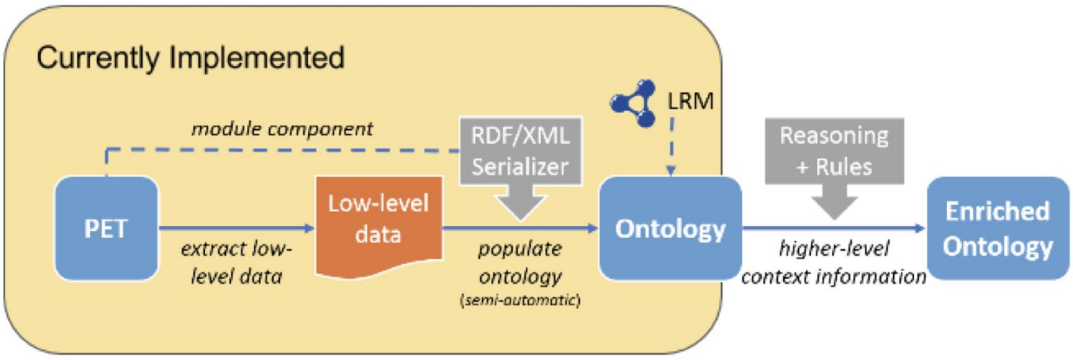


Fig. 5-7. Status and planned advances in the analysis of environment information.

In this specific task we have worked on extending the PET2LRM (**Version2**) to support the export of event specific data from PET; this use context information data is used in the task for analysis. Fig. 5-7 illustrates the current status of PET2LRM along with the features to be integrated in the near future. For a more detailed description of the PET2LRM, we refer to D4.2 [PERICLES D4.2, 2015]; and to D4.1 [PERICLES D4.1, 2014] for the description of the PET tool, and its approach and data model.

In future tasks T4.4 (due M40) and T4.5 (due M44), we will report the work on the PET2Ecosystem implementation and the derivation of higher level information from PET-ecosystem data, such as higher level dependencies or dependency chains and other ecosystem entities, which could more efficiently connect the PET-extracted information to the rest of the ecosystem graph.

By embedding into a DO its use context information from PET (e.g. when a file was opened, who opened the file, which software is used for accessing the DO, etc.), we will support the tracking of the use of the data and its spread in communities; and also the inference of higher level information such as information dependencies (dependencies between information entities) also taking into account user communities.

5.2.1. Improvements to the PET2LRM Use Context Information Export

Since its inception, the PET framework has been created with the aim to collect, among other, a wide range of information from the environment of use of Digital Objects. In the scope of this task, we have focused on extending PET2LRM (***Version2***) so that the event data generated by the modules can be exported in LRM form; together with the improvement of the LRM output. This can enable the collection of events in form of provenance information that can be used to infer the structure of casual, not formally described workflows, as already announced in PERICLES D4.1 [PERICLES D4.1, 2014], such as the case where a scientist is executing different tests to experiment with different versions of an algorithm; or applying a sequence of different processing steps manually.

The tracking of this use context data can be valuable both for provenance information (creation of dependencies between data and software versions), as for allowing the discovery of the data transformation workflows that are themselves often not formalised.

The PET2LRM code, executable, and sample data used in this task can be accessed in the internal PERICLES repository and will receive further improvements and extension to the DEM model in future WP4 tasks.

In the scope of this task, some exemplar data was also generated to cover the different type of information extracted in PET: environment, file specific, and monitoring (events) output; and some sample data for the use case of provenance information described in this paragraph.

The PET2LRM tool can be executed in a command shell with the following syntax:

```
java -jar pet2lrm.jar -d "PATH"
```

where "PATH" points to the main PET folder where the PET data is stored. The PET tool itself is not required and the tool can be run using only the PET data folder. The default output file is called "sei_model.owl" and contains all the data that is mapped to the LRM ontology.

In Fig. 5-8 we are showing an excerpt of the PET2LRM output, related to event and provenance information. Further analysis of the use of this data is described in the next paragraph.

```
<rdf:Description
rdf:about="http://www.pericles-project.eu/ns/PET2LRM#1448447761833-LSOF
_use_monitor">
    <pet2lrm:command>Microsoft</pet2lrm:command>
    <rdf:type
rdf:resource="http://www.pericles-project.eu/ns/PET2LRM#MonitoringEvent
"/>
    <pet2lrm:timestamp>1448447761833</pet2lrm:timestamp>
    <pet2lrm:fullProcessName>Microsoft Word</pet2lrm:fullProcessName>
    <pet2lrm:type>appear</pet2lrm:type>
    <pet2lrm:device>1,8</pet2lrm:device>
    <lrm:realizes
rdf:resource="http://www.pericles-project.eu/ns/PET2LRM#MonitoringEvent
"/>
    <pet2lrm:isExtractedBy>LSOF use monitor</pet2lrm:isExtractedBy>
    <pet2lrm:processArguments>[/Applications/Microsoft Office
2011/Microsoft Word.app/Contents/MacOS/Microsoft
Word]</pet2lrm:processArguments>
    <pet2lrm:hasFileName>/Users/fabio/Downloads/PET/test
doc.doc</pet2lrm:hasFileName>
    <pet2lrm:dateTime
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2015-11-25T11:
36:01+01</pet2lrm:dateTime>
    <pet2lrm:size>22016</pet2lrm:size>
    <pet2lrm:pid>7410</pet2lrm:pid>
    <pet2lrm:fullProcessPath>/Applications/Microsoft Office
2011/Microsoft Word.app/Contents/MacOS/Microsoft
Word</pet2lrm:fullProcessPath>
</rdf:Description>
[...]
```

Fig. 5-8. Sample of PET2LRM event data of use context information.

5.2.2. Analysis of Use Context from PET2LRM

As previously mentioned, the PET tool can be used to capture the digital Environment where DO exist, and to formally describe it according to the notions introduced in PET2LRM model. Such information may be related to:

- The **system** itself (sei_model_environment.owl output) - technical system information, software, hardware and data installations, etc.
- The analysed/stored **files** (sei_model_files.owl output) - such as metadata, file type, checksum, author, access dates, xml output from MediaInfo[27] (when applicable), etc.
- The **events** monitored/captured within the system (sei_model_monitoring.owl output) - such as when was a file opened, who opened the file, which software was used for accessing the file, where is it located, which is the size of it, etc.

More specifically, in the extraction of ***sei_model_environment.owl* file** from PET2LRM, an instance of `pet2lrm:AbstractEnvironment` class may be populated to encapsulate system information through relations and relevant instantiations attached to it; relevant object properties of PET2LRM model are the following: *hasMemory*, *hasGraphicCard*, *hasCPU*, *hasOperatingSystem*, *hasFile*, *hasJavaInstallation*, etc. Such relations can be transformed into representations of the use context of the analysed environment, with the use of the notion `lrm:Dependency` and its involved properties, as well as with their extensions, as introduced in the A&M domain. For example, instances attached to the property `pet2lrm:hasOperatingSystem` may populate a (software) dependency with functional intention, while, instances involved directly/indirectly in a `pet2lrm:hasMemory` may instantiate a (hardware) dependency from (`lrm:from`) a resource or an environment to (`lrm:to`) a specific instantiation of computer memory. Proper SPARQL queries can be formed and executed so as to retrieve all the available information needed and to further perform the corresponding dependency instantiations, as those given in Table 5-1.

Table 5-1. SPARQL queries to generate instantiations of Software and Hardware Dependencies derived from PET2LRM output representation.

Software Dependency	SPARQL SELECT (from PET2LRM output)

	PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
	PREFIX pet2lrm: <http://www.pericles-project.eu/ns/PET2LRM#>
	
	SELECT **?abstractEnvironment ?operatingSystemName ?operatingSystemInstance**
	WHERE
	{
	?abstractEnvironment **pet2lrm**:hasOperatingSystem **?operatingSystem** .
	
	?operatingSystem **rdf:**type **pet2lrm**:AbstractExtractionResult .
	?operatingSystem **pet2lrm**:operatingSystem **?operatingSystemInstance** .
	
	?operatingSystemInstance **rdf:**type **pet2lrm**:OperatingSystem .
	?operatingSystemInstance **pet2lrm**:os_name **?operatingSystemName** .
	}

	SPARQL UPDATE (insert previously selected instances to domain ontology)
	PREFIX anm: <http://temporary_link/ArtAndMedia#> PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

[27] https://mediaarea.net/en/MediaInfo

PREFIX lrm: <http://xrce.xerox.com/LRM#>

INSERT DATA
{

?abstractEnvironment	**rdf**:type	**pet2lrm**:AbstractEnvironment .
?softwareDependency	**rdf**:type	**anm**:SoftwareDependency .
?operatingSystemInstance	**rdf**:type	**anm**:OperatingSystem .
?operatingSystemInstance	**rdfs**:label	'operating_system_name' .
?softwareDependency	**lrm**:from	**?abstractEnvironment** .
?softwareDependency	**lrm**:to	**?operatingSystemInstance** .
?softwareDependency	**lrm**:intention	'functional' .
?softwareDependency	**lrm**:specification	'.....' .

}

Hardware Dependency

SPARQL SELECT (from PET2LRM output)

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pet2lrm: <http://www.pericles-project.eu/ns/PET2LRM#>

SELECT **?abstractEnvironment ?memoryInstance ?totalMemoryValue ?actualFreeMemoryValue ?actualUsedMemoryValue**
WHERE
{

?abstractEnvironment	**pet2lrm**:hasMemory	**?memory** .
?memory	**rdf:**type	**pet2lrm**:AbstractExtractionResult .
?memory	**pet2lrm**:memory	**?memoryInstance** .
?memoryInstance	**rdf:**type	**pet2lrm**:Memory .
?memoryInstance	**pet2lrm**:total	**?totalMemoryValue** .
?memoryInstance	**pet2lrm**:actualFree	**?actualFreeMemoryValue** .
?memoryInstance	**pet2lrm**:actualUsed	**?actualUsedMemoryValue** .

}

SPARQL UPDATE (insert previously selected instances to domain ontology)

PREFIX anm: <http://temporary_link/ArtAndMedia#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lrm: <http://xrce.xerox.com/LRM#>

INSERT DATA
{

?computer	**rdf**:type	**anm**:Computer .
?memoryInstance	**rdf**:type	**anm**:Memory .
?memoryInstance	**anm**:hasSize	**?memoryInstanceSize** .

```
        ?memoryInstanceSize     anm:hasValue      'total_memory_value' .

        ?hardwareDependency     rdf:type          anm:HardwareDependency .
        ?hardwareDependency     lrm:from          ?computer .
        ?hardwareDependency     lrm:to            ?memoryInstance .
        ?hardwareDependency     lrm:intention     'functional' .
        ?hardwareDependency     lrm:specification '.....' .
}
```

The extraction of ***sei_model_files.owl*** **file** from PET2LRM tool records metadata and general information about files, stored in the environment (computer) analysed. Focusing on the metadata extracted by the *Calculate file checksum* PET2LRM module (see Fig. 5-9) we present all the detailed information regarding the checksum data attached to a specific file.

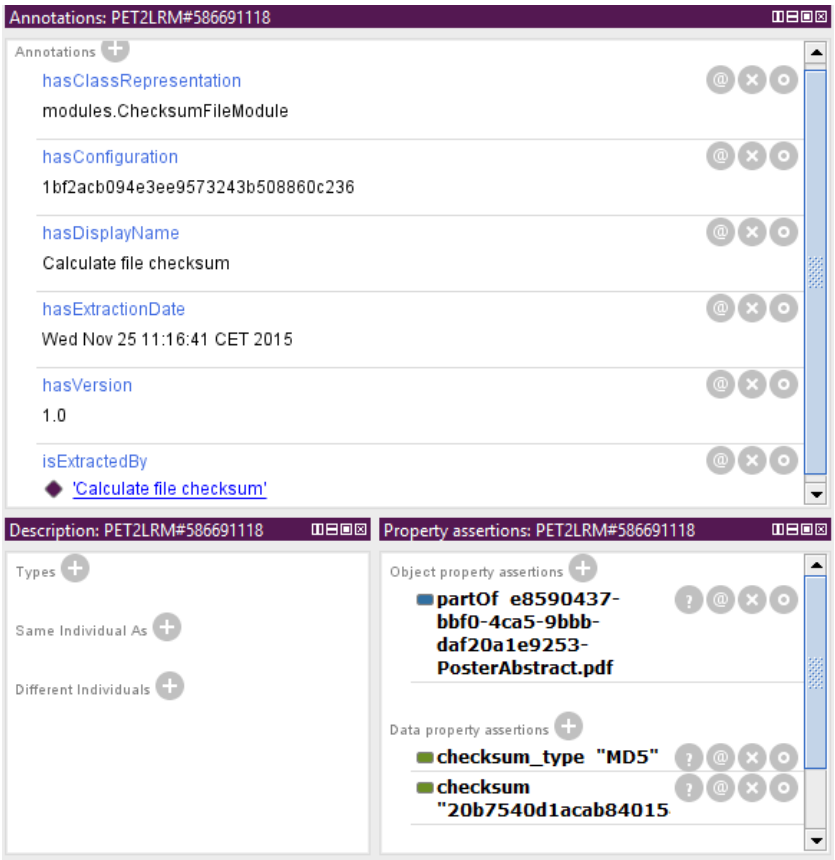


Fig. 5-9. Instantiation of properties for a specific file, with information extracted from *Calculate file checksum* PET2LRM module[28].

The above metadata collected by PET together with the use data, describe a dependency relationship between the digital file and some significant knowledge: such an example could be transformed into a Data Dependency from a digital file to the knowledge of specific data (checksum), with functional intention. This dependency may describe the fact that, in order to assure the file integrity, the user or the system needs to know the checksum validity algorithm and value. The described dependency can be populated by performing proper SPARQL queries, as seen in Table 5-2.

[28] Snapshot taken from Protégé (http://protege.stanford.edu/).

Table 5-2. SPARQL queries to generate instantiations of Data Dependency derived from PET2LRM output representation.

Data Dependency	SPARQL SELECT (from PET2LRM output)

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pet2lrm: <http://www.pericles-project.eu/ns/PET2LRM#>
PREFIX lrm: <http://xrce.xerox.com/LRM#>

SELECT ?calculateFileChecksum ?checksumType ?checksum ?fileName
WHERE
{
        ?calculateFileChecksum  pet2lrm:checksum        ?checksum .
        ?calculateFileChecksum  pet2lrm:checksum_type   ?checksumType .
        ?calculateFileChecksum  lrm:partOf              ?file .

        ?file                   rdf:type                pet2lrm:AbstractPart .
        ?file                   pet2lrm:hasFileName     ?fileName .
}
```

SPARQL UPDATE (insert previously selected instances to domain ontology)

```
PREFIX anm: <http://temporary_link/ArtAndMedia#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lrm: <http://xrce.xerox.com/LRM#>

INSERT DATA
{
        ?file                   rdf:type                anm:DigitalFile .
        ?file                   rdfs:label              'file_name' .

        ?checksum               rdf:type                anm:Checksum .
        ?checksum               anm:hasChecksumType     'checksum_type' .
        ?checksum               anm:hasChecksumString   'checksum' .

        ?dataDependency         rdf:type                anm:DataDependency .
        ?dataDependency         lrm:from                ?file .
        ?dataDependency         lrm:to                  ?checksum .
        ?dataDependency         lrm:intention           'functional' .
        ?dataDependency         lrm:specification       '.....' .
}
```

The extraction of ***sei_model_monitoring.owl*** **file** from PET tool keeps event information as they were triggered and tracked within the analysed system. An example event could be the access (opening) of a file by a user; an indicative snapshot of properties instantiated in the corresponding PET2LRM class (`pet2lrm:MonitoringEvent`) when the event was performed, is presented in Fig. 5-10.

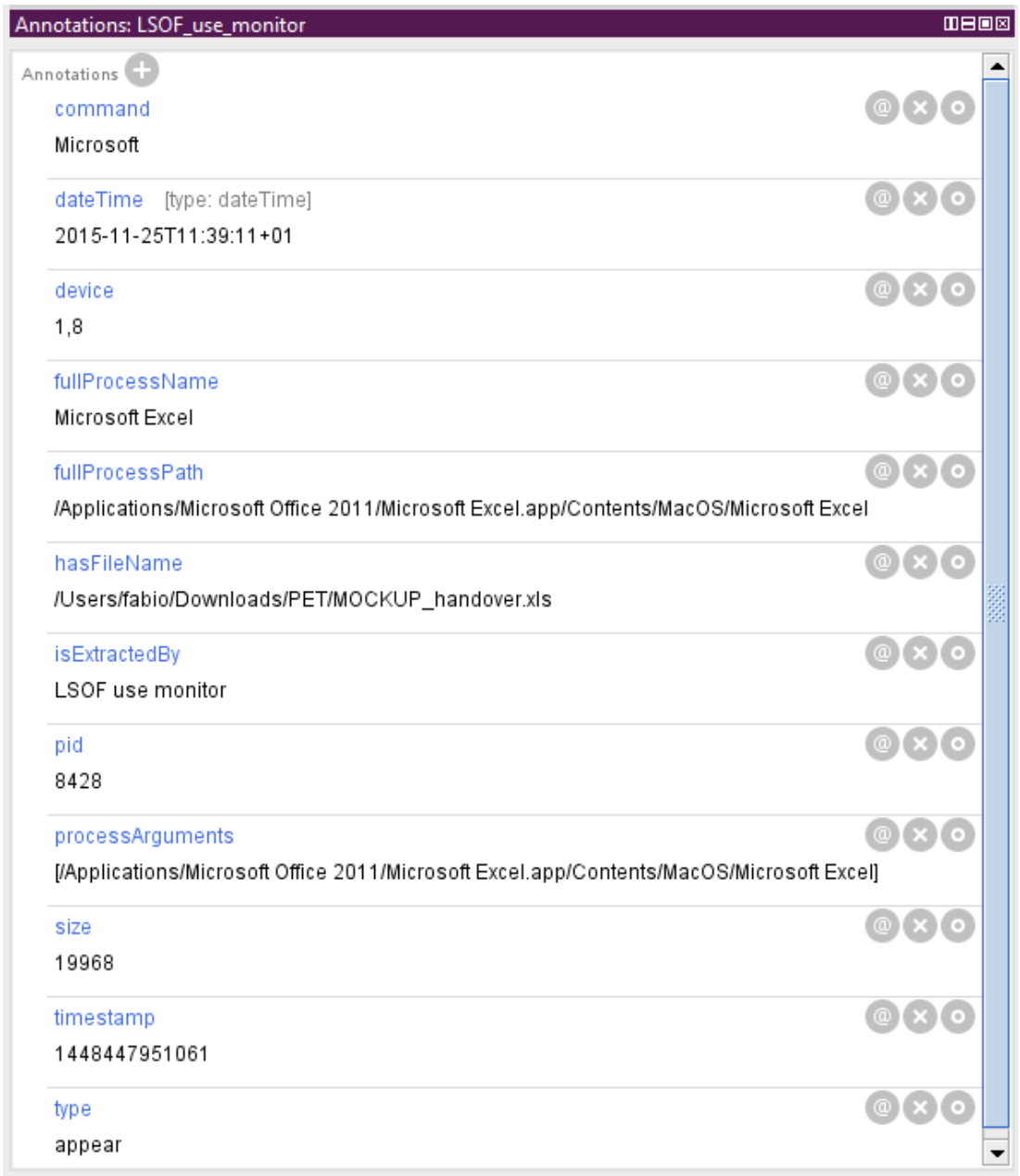


Fig. 5-10. Instantiation of `pet2lrm:MonitoringEvent` class' properties, while opening a specific digital file[29].

Information such as the above could be useful to create instantiations in the A&M ontology that illustrate the corresponding use context of the described event. In the scenario presented in the previous figure, it is referenced that the example file (*MOCUP_handover.xls*) has been successfully accessed with the use of a specific software type and version (*Microsoft Excel*, *2011*). Based on the aforementioned, a software dependency may be populated so as to assess the fact that there is a demand for a specific software application (surely, not the only one) in order to open the file successfully. The described dependency can be populated by performing proper SPARQL queries, as seen in Table 5-3, while an illustration of the described dependency is given in Fig. 5-11.

[29] Snapshot taken from Protégé (http://protege.stanford.edu/).

Table 5-3. SPARQL queries to generate instantiations of Software Dependency derived from PET2LRM output representation.

Software Dependency

SPARQL SELECT (from PET2LRM output)

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pet2lrm: <http://www.pericles-project.eu/ns/PET2LRM#>
PREFIX lrm: <http://xrce.xerox.com/LRM#>

SELECT ?monitoringEvent ?fileName ?computerProgram
WHERE
{
        ?monitoringEvent    rdf:type                pet2lrm:MonitoringEvent .
        ?monitoringEvent    pet2lrm:fullProcessName ?computerProgram.
        ?monitoringEvent    pet2lrm:hasFileName     ?fileName .
}
```

SPARQL UPDATE (insert previously selected instances to domain ontology)

```
PREFIX anm: <http://temporary_link/ArtAndMedia#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lrm: <http://xrce.xerox.com/LRM#>

INSERT DATA
{
        ?file               rdf:type                anm:DigitalFile .
        ?file               rdfs:label              'file_name' .

        ?computerProgram    rdf:type                anm:ComputerProgram .
        ?computerProgram    rdfs:label              'computer_program_name' .

        ?checksum           anm:hasVersion          'version_value' .

        ?softwareDependency rdf:type                anm:SoftwareDependency .
        ?softwareDependency lrm:from                ?file .
        ?softwareDependency lrm:to                  ?computerProgram .
        ?softwareDependency lrm:intention           'functional' .
        ?softwareDependency lrm:specification       '.....' .
}
```

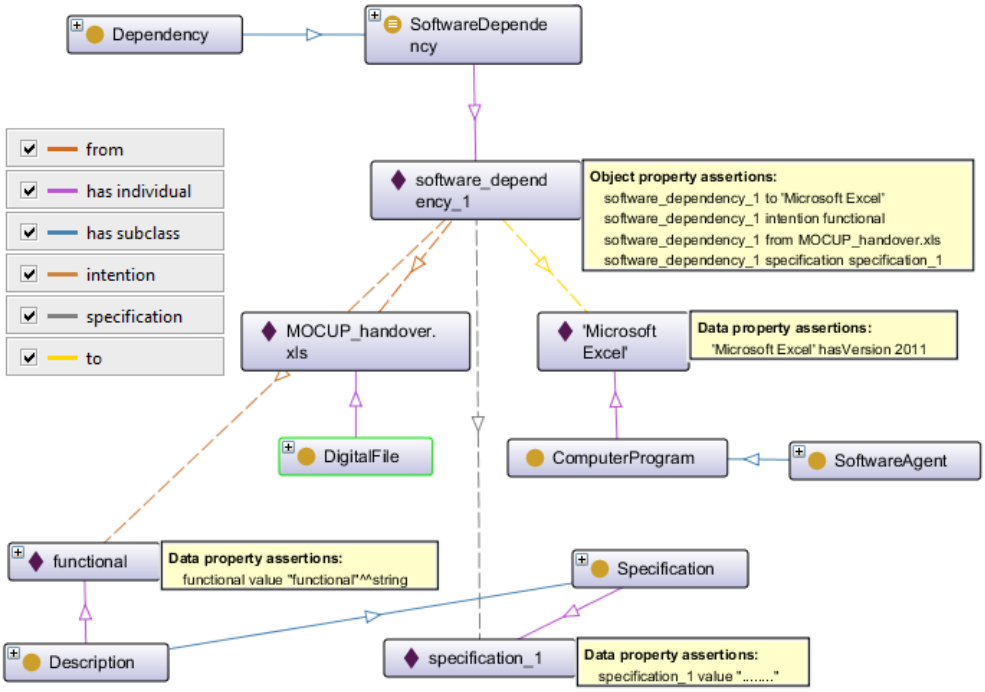


Fig. 5-11. A Software Dependency for a digital file that was accessed by a specific software[30].

A more sophisticated scenario could be the following: a python script file is stored and executed in a computer, in order to calibrate raw data. The PET tool is capable of capturing (a) the software version (interpreter) used to execute the script file, (b) the execution parameters of the execution command, (c) the content of the script itself, and (d) the output of the execution. Next, the PET2LRM populates instances of `pet2lrm:MonitoringEvent` class with information derived from PET tool. In our implemented case, we focus in information regarding the software (or library) and the execution arguments. Such a scenario would be handled with a software and a data dependency created by SPARQL queries as shown in Table 5-4.

Table 5-4. SPARQL queries to generate instantiations of Software and Data Dependencies derived from PET2LRM output representation.

	SPARQL SELECT (from PET2LRM output)

Dependency	PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
	PREFIX pet2lrm: <http://www.pericles-project.eu/ns/PET2LRM#>
	
	SELECT **?monitoringEvent ?processName ?executionArguments ?fileName**
	WHERE
	{
	?monitoringEvent **rdf:**type **pet2lrm**:MonitoringEvent.
	?monitoringEvent **pet2lrm**:fullProcessName **?processName** .
	?monitoringEvent **pet2lrm**:processArguments **?executionArguments** .
	?monitoringEvent **pet2lrm**:hasFileName **?fileName** .
	}

[30] Snapshot created with Protégé plugin Ontograf (http://protegewiki.stanford.edu/wiki/Ontograf).

SPARQL UPDATE (insert previously selected instances to domain ontology)

```
PREFIX anm: <http://temporary_link/ArtAndMedia#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lrm: <http://xrce.xerox.com/LRM#>

INSERT DATA
{
        ?monitoringEvent        rdf:type        anm:ExecutionActivity .
        ?processName            rdf:type        lrm:SoftwareAgent .
        ?executionArguments     rdf:type        anm:Data .
        ?fileName               rdf:type        anm:DigitalFile .

        ?softwareDependency     rdf:type        anm:SoftwareDependency .
        ?dataDependency         rdf:type        anm:DataDependency .

        ?softwareDependency     lrm:from        ?monitoringEvent .
        ?softwareDependency     lrm:to          ?processName .
        ?softwareDependency     lrm:intention   'functional' .
        ?softwareDependency     lrm:specification '.....' .

        ?dataDependency         lrm:from        ?monitoringEvent .
        ?dataDependency         lrm:to          ?executionArguments .
        ?dataDependency         lrm:to          ?fileName .
        ?dataDependency         lrm:intention   'functional' .
        ?dataDependency         lrm:specification '.....' .
}
```

An example of generated instances from the aforementioned queries is presented in Fig. 5-12.

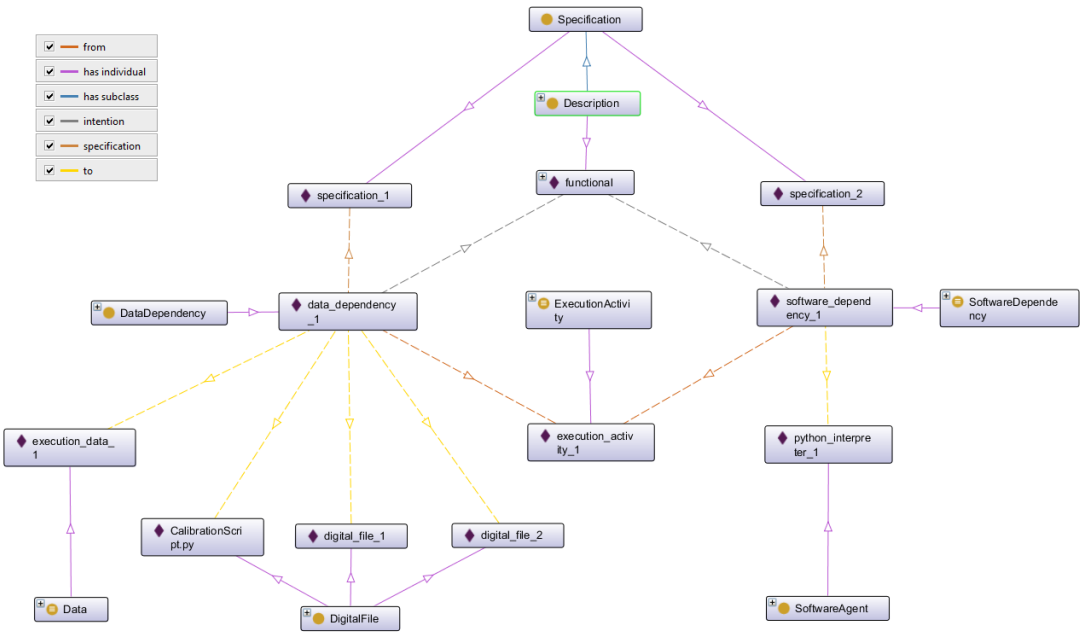


Fig. 5-12. Software and Hardware Dependencies' instantiations for the execution activity of a script file that uses a Python interpreter.

5.3. Chapter Summary

Chapter 5 demonstrates the approach of **extraction and analysis of use context** based on the ontology-based representations within PERICLES. The model that has been followed to formally represent the use context of Digital Objects (DO) that interact within a Digital Environment, is described in detail in Section 5.1; entities like `lrm:Agent`, `lrm:Activity`, `lrm:Resource` and `lrm:Dependency`, as well as their extensions, are instantiated and combined accordingly, so as to describe, in terms of ontologies, the context of use, existence and reliance of DO.

Furthermore, various tools and mechanisms developed within PERICLES have been successfully combined in order to **produce practical representations of the use context** (see Section 5.2): the PET tool can be utilised to extract useful information from the environment of execution and use of DO; like for example software, hardware and data installations of a computer as well as events triggered when manipulating digital files. Formalisation of PET output is achieved with the use of PET2LRM tool, while proper analysis of PET2LRM output and mechanisms for mapping it into use context related formal representations (see Section 5.2.2) have been established successfully.

6. Conclusions and Next Steps

6.1. Conclusions

This deliverable reported on the work conducted in T4.3, focusing on the extraction and analysis of semantic information from visual and textual digital content, as well as on the retrieval and analysis of the use context of digital objects. More specifically, the following outputs per topic were presented:

- **Content decomposition and feature extraction**: For analysing visual content, the deliverable presented **a pipeline approach for image representation**, consisting of the following steps: (a) key-point detection, (b) feature extraction, and (c) feature encoding. Because this approach leads to massive data, data compression and reconstruction during the extraction of semantic information are of uttermost importance. In this direction, we proposed **GDS**, **a novel approach for measuring sparsity** that extends the limits of the adopted methodologies for compression. Regarding text-based content, the chapter introduced the **field theory of semantic content**, a novel integrative model based on established theories of word meaning from theoretical linguistics to model the evolution of document content over time. A set of experimental results validating the proposed approach were also presented.
- **Semantic concept detection and content classification**: The deliverable underpinned the importance of semantic information extraction and **proposed prototype methods for extracting semantic information from visual and text-based content**. More specifically, the document presented: (a) **SALIC**, a novel method for detecting semantic concepts from the content of a DO that automatically gathers training data without requiring significant annotation efforts, while at the same time minimizing the number of the required training instances and increasing the performance of the classification models by utilizing a smart sampling approach, (b) **PCS**, a novel and very fast method for dimensionality reduction. Further we introduced a tool for the scalable processing of semantic media content based on vector fields which paves the way for the exploration of quantum-like content behaviour in digital collections, including the typology of correlations underlying machine learning for automatic content analysis. With regards to semantic information extraction from text-based content, the deliverable also presented our approaches for **analysing source text documents** relevant to the two case studies and using the extracted information for **populating the developed domain ontologies with instances**.
- **Extraction and analysis of use context information**: Finally, the deliverable presented our proposed approaches for (a) **representing use context** (i.e. information related to contexts of use of the DO), and, (b) **extracting and analysing use context information**. For the former we deployed the domain ontologies developed within WP2 and, more specifically, the "Dependency" construct and its enclosed notions of "Intention" and "Specification", along with the specialisations of these concepts. For the latter we used two WP4 outputs, i.e. the PET tool along with its PET2LRM plugin that allows converting PET output into LRM fragments.

The proposed tools and algorithms are designed to be flexible, meaning that they can be used standalone, but also in combination with test scenarios from WP6.

6.2. Next Steps

Future work can be separated in two streams, firstly the follow-up on activities started in T4.3 for the upcoming PERICLES tasks (like, e.g., T4.4 and T4.5), and secondly, those beyond the lifetime of PERICLES, securing the sustainability of our research findings and considerations.

Links to upcoming PERICLES tasks

With regard to the major topics in this deliverable, namely content decomposition and feature extraction, semantic concept detection and content classification, and extraction and analysis of use context information, for T4.4 these research tracks will be integrated into a single experimental workflow where, using both vector space and probabilistic methods for contextually dependent semantic content analysis on image and text data and metadata, we will test the tools presented above for several kinds of socially induced concept and topic shifts. This experimental workflow will exemplify DP considerations in WP5 and processing challenges in WP6.

Regarding the latter WP, we have already started to closely collaborate with WP6 in order to investigate whether the specified user scenarios can be adequately supported by the current versions of the D4.3 outputs, or whether additional updates and modifications will be necessary. However, since the tools are generic, we are aiming to also investigate whether more abstract scenarios could be successfully deployed.

The workhorse for the planned series of experiments will be publicly available catalog metadata from Tate which holds both image and text descriptors, hopefully suitable to bridge the semantic gap between running text-based vs. image-based topic interpretation, while at the same time representing the lower boundary of scalable data.

On the other hand, for T4.5, we plan to pursue two major tracks. One will be looking into the mathematical foundations of correlations exploited by machine learning, plus an analytical effort to learn about the applicability of classical vs. quantum mechanics - or any alternatives between these two extremes - to represent evolving semantic content; the other will explain the match between evolving ontologies and evolving vector spaces used for automatic reasoning. As both tracks will be highly multidisciplinary and novel, we expect both negative and positive results, indicating possible dead end streets as well as directions to be explored.

Links to beyond PERICLES

All research activities in WP4 are unfinished by their innovative nature, therefore several different tracks of inquiries can be expected to continue beyond PERICLES. To mention but a few, one major track will be to work out our recently proposed “Biology Metaphor” [PERICLES D4.1, 2014], e.g. by clarifying the relationship between significant properties as part of the “genome” of DOs and their linkage with real biological data from living organisms. Another track will doubtlessly focus on semantic content as it manifests itself in different digital media, including its evolution, modeling, and reuse. A third focus should be to work out the nature and extent of parallels between language-based DO representations and their “behaviour” over time, i.e. the limits of physics as a metaphor for socially driven semantic content. As a sidetrack, one could even sum up the above in terms of an upcoming new vision focusing on “information cosmology”, modelling expanding semantic content into disciplinary and genre-specific discourse universes. Finally, ongoing scalability and tool development efforts will translate the above to, and ferment the future development of, LTDP.

7. References

[Abel et al., 2012] Abel, F., Hauff, C., Houben, G. J., Stronkman, R., & Tao, K. (2012, June). Semantics+ filtering+search=twitcident. Exploring information in social web streams. *Proc. 23rd ACM Conf. on Hypertext and Social Media*, pp. 285-294. ACM.

[Aerts & Gabora, 2005] Aerts, D. and Gabora, L. (2005). A theory of concepts and their combinations I: The structure of the sets of contexts and properties. *Kybernetes*, 34(1/2):151–175.

[Aharon et al., 2006] Aharon, M., Elad, M., & Bruckstein, A. (2006). K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *Signal Processing, IEEE Transactions on*, 54(11), 4311-4322.

[Atkinson & Flint, 2001] Atkinson, R., & Flint, J. (2001). Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social research update*, 33(1), 1-4.

[Bacardit & Llorà, 2013] Bacardit, J., & Llorà, X. (2013). Large-scale data mining using genetics-based machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 37-61.

[Bai et al., 2007] Bai, J., Nie, J. Y., Cao, G., & Bouchard, H. (2007, July). Using query contexts in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 15-22). ACM.

[Baraniuk, 2007] Baraniuk, R. G. (2007). Compressive sensing. *IEEE signal processing magazine*, 24(4).

[Baraniuk et al., 2011] Baraniuk, R., Davenport, M. A., Duarte, M. F., & Hegde, C. (2011). *An introduction to compressive sensing*. Connexions e-textbook.

[Bawden et al., 2015] Bawden, D., Robinson, L. and Siddiqui, T. (2015). “Potentialities or possibilities”: Towards quantum information science? *Journal of the Association for Information Science and Technology*, 66, 3, 437–449.

[Bay et al., 2008] Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer vision and image understanding*, 110(3), 346-359.

[Becker et al., 2011] Becker, H., Naaman, M., & Gravano, L. (2011). Beyond Trending Topics: Real-World Event Identification on Twitter. *ICWSM*, 11, 438-441.

[Biernacki & Waldorf, 1981] Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research*, 10(2), 141-163.

[Bizer et al., 2009] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S. (2009). DBPedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 7, Issue 3 (pp. 154-165). Elsevier.

[Bober, 2001] Bober, M., (2001) "MPEG-7 visual shape descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.11, no.6, pp.716,719

[Breiman et al., 1984] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

[Brodtkorb et al., 2010] Brodtkorb, A., Dyken, C., Hagen, T., Hjelmervik, J., and Storaasli, O. (2010). State-of-the-art in heterogeneous computing. *Scientific Programming*, 18(1):1–33.

[Bruner, 1961] Bruner, J. S. (1961). *The act of discovery*. Harvard educational review.

[Byna et al., 2010] Byna, S., Meng, J., Raghunathan, A., Chakradhar, S., and Cadambi, S. (2010). Best-effort semantic document search on GPUs. *Proc. 3rd Workshop on General-Purpose Computation on Graphics Processing Units (GPGPU-10)*, pp. 86–93, New York, NY, USA.

[Cai et al., 2007] Cai, K., Chen, C., Bu, J., Huang, P., & Kang, Z. (2007, May). Exploration of query context for information retrieval. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1157-1158). ACM.

[Cai et al., 2011] Cai, H., Mikolajczyk, K., & Matas, J. (2011). Learning linear discriminant projections for dimensionality reduction of image descriptors. *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, 33(2), 338-352.

[Candès & Tao, 2005] Candes, E. J., & Tao, T. (2005). Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12), 4203-4215.

[Candès & Tao, 2007] Candès, E., & Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313-2351.

[Candès, 2008] Candès, E. J., Wakin, M. B., & Boyd, S. P. (2008). Enhancing sparsity by reweighted ℓ 1 minimization. *Journal of Fourier analysis and applications*, 14(5-6), 877-905.

[Candès & Walkin, 2008] Candès, E. J., & Wakin, M. B. (2008). An introduction to compressive sampling. *Signal Processing Magazine*, IEEE, 25(2), 21-30.

[Cavanagh et al., 2009] Cavanagh, J., Potok, T., and Cui, X. (2009). Parallel latent semantic analysis using a graphics processing unit. *Proc. 11th Annual Conf. Companion on Genetic and Evolutionary Computation Conference (GECCO-09): Late Breaking Papers*, pp. 2505-2510, Montreal, Canada.

[Cevher et al., 2008] Cevher, V., Sankaranarayanan, A., Duarte, M. F., Reddy, D., Baraniuk, R. G., & Chellappa, R. (2008). Compressive sensing for background subtraction. *Computer Vision–ECCV 2008* (pp. 155-168). Springer Berlin Heidelberg.

[Chang et al., 2005] Chang, E., Tong, S., Goh, K., & Chang, C.-W. (2005). Support Vector Machine Concept-Dependent Active Learning For Image Retrieval. *IEEE Transactions on Multimedia*.

[Chang & Lin, 2011] Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.

[Chatfield et al., 2011] Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods, *British Machine Vision Conf.*

[Chatfield et al., 2014] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets, *British Machine Vision Conf.*

[Chatzilari et al., 2014] Chatzilari, E., Nikolopoulos, S., Kompatsiaris, Y., & Kittler, J. (2014). Active learning in social context for image classification. *9th Int. Conf. on Computer Vision Theory and Applications, VISAPP*.

[Chatzilari et al., 2015] Elisavet Chatzilari, Spiros Nikolopoulos, Yiannis Kompatsiaris, Josef Kittler, "SALIC: Social Active Leaning for Image Classification, *IEEE Transactions on Multimedia*, under review.

[Chen et al., 2003] Chen, H., Finin, T., & Joshi, A. (2003). An ontology for context-aware pervasive computing environments. *The Knowledge Engineering Review*, 18(03), 197-207.

[Chennubhotla & Jepson, 2001] Chennubhotla, C., & Jepson, A. (2001). Sparse PCA. Extracting multi-scale structure from data. *Proc. 8th IEEE Int. Conf. on Computer Vision (ICCV 2001)*, Vol. 1, pp. 641-647.

[Chua et al., 2009] Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009, July). NUS-WIDE: a real-world web image database from National University of Singapore. *Proc. ACM Int. Conf. on Image and Video Retrieval* (p. 48). ACM.

[Cohn et al., 1994] Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine learning*, 15(2), 201-221.

[Cool & Spink, 2002] Cool, C., & Spink, A. (2002). Issues of context in information retrieval (IR): an introduction to the special issue. *Information Processing & Management*, 38(5), 605-611.

[Cortes & Vapnik, 1995] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

[Cotter et al., 2013] Cotter, A., Shalev-Shwartz, S., & Srebro, N. (2013). Learning optimally sparse support vector machines. *Proc. 30th Int. Conf. on Machine Learning (ICML-13)*, pp. 266-274.

[Cox & Cox, 1994] Cox, T. and Cox, M. (1994). *Multidimensional Scaling*. Chapman and Hall.

[Dalton, 1920] Dalton, H. (1920). The measurement of the inequality of incomes. *The Economic Journal*, 348-361.

[Dappert & Farquhar, 2009] Dappert, A., & Farquhar, A. (2009). Significance is in the eye of the stakeholder. *Research and Advanced Technology for Digital Libraries* (pp. 297-308). Springer Berlin Heidelberg.

[Dappert & Enders, 2010] Dappert, A. & Enders, M. (2010). Digital preservation metadata standards. Information Standards Quarterly, 22 (2) 5-13. Available at: https://goo.gl/Bl4pjt.

[Darányi & Wittek, 2012] Darányi, S. and Wittek, P. (2012). Connecting the dots: Mass, energy, word meaning, and particle-wave duality. *Proc. 6th Int. Quantum Interaction Symp. (QI-12)*, 207–217.

[Daróczy et al., 2011] Daróczy, B., Pethes, R., and Benczúr, A. (2011). SZTAKI @ ImageCLEF 2011. *Proc. Conf. on Multilingual and Multimodal Information Access Evaluation (CLEF-11)*, Amsterdam, The Netherlands.

[Davenport et al., 2011] Davenport, M. A., Duarte, M. F., Eldar, Y. C., & Kutyniok, G. (2011). *Introduction to compressed sensing*. Preprint, 93, 1-64.

[Dean & Ghemawat, 2008] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113.

[Deng et al., 2009a] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 248-255. IEEE.

[Deng et al., 2009b] Deng, H., King, I., & Lyu, M. R. (2009, November). Enhancing expertise retrieval using community-aware strategies. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1733-1736). ACM.

[Derntl & Hummel, 2005] Derntl, M., & Hummel, K. (2005, March). Modeling context-aware e-learning scenarios. *Pervasive Computing and Communications Workshops, 2005. PerCom 2005 Workshops. Third IEEE International Conference on* (pp. 337-342). IEEE.

[Dey et al., 2001] Dey, A. K., Abowd, G. D., & Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-computer interaction*, 16(2), 97-166.

[Dichev & Dicheva, 2005] Dichev, C., & Dicheva, D. (2005, October). Context as collection of alternatives. *Proceedings of the International Semantic Web for E-Learning Workshop at the 3rd International Conference on Knowledge Capture* (pp. 53-58).

[Ding et al., 2009] Ding, S., He, J., Yan, H., and Suel, T. (2009). Using graphics processors for high performance IR query processing. *Proc. 18th Int. Conf. on World Wide Web (WWW-09)*, pages 421–430, Spain, Madrid.

[Domingos & Pazzani, 1997] Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, 29(2-3), 103-130.

[Donoho et al., 2006] Donoho, D. L., Elad, M., & Temlyakov, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1), 6-18.

[Donoho & Tsaig, 2008] Donoho, D. L., & Tsaig, Y. (2008). Fast solution of-norm minimization problems when the solution may be sparse. *Information Theory, IEEE Transactions on*, 54(11), 4789-4812.

[Dotan & Zaphiris, 2010] Dotan, A., & Zaphiris, P. (2010). A cross-cultural analysis of Flickr users from Peru, Israel, Iran, Taiwan and the UK. *Int. Journal of Web Based Communities*, 6(3), 284-302.

[Ebert et al., 2012] Ebert, S., Fritz, M., & Schiele, B. (2012, June). Ralf: A reinforced active learning formulation for object class recognition. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3626-3633. IEEE.

[Egmont-Petersen et al., 2002] Egmont-Petersen, M., de Ridder, D., & Handels, H. (2002). Image processing with neural networks—a review. *Pattern recognition*, 35(10), 2279-2301.

[Ejigu et al., 2007] Ejigu, D., Scuturici, M., & Brunie, L. (2007, March). An ontology-based approach to context modeling and reasoning in pervasive computing. *Pervasive Computing and Communications Workshops, 2007. PerCom Workshops' 07. Fifth Annual IEEE International Conference on* (pp. 14-19). IEEE.

[Elad & Aharon, 2006] Elad, M., & Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12), 3736-3745.

[Everingham et al., 2007] Everingham, M., Gool, L., Williams, C.-K., Winn, J., & Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[Everingham et al., 2012] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2012). The PASCAL visual object classes challenge 2012 (VOC2012) results.

[Falcão, 2010] Falcão, P. (2010), Developing a Risk Assessment Tool for the conservation of software-based artworks. MA-Thesis, Swiss Conservation - Restoration Campus.

[Fang et al., 2014] Fang, M., Yin, J., & Tao, D. (2014, June). Active learning for crowdsourcing using knowledge transfer. *Proc. 28th AAAI Conf. on Artificial Intelligence*.

[Fellbaum, 1998] Fellbaum, W. (1998). *An Electronic Lexical Database (Language, Speech, and Communication)*.

[Firth, 1957] Firth, J.R. 1957. *Papers in Linguistics 1934–1951*. London: Oxford University Press.

[Freytag et al., 2014] Freytag, A., Rodner, E., & Denzler, J. (2014). Selecting Influential Examples: Active Learning with Expected Model Output Changes. *Computer Vision–ECCV 2014* (pp. 562-577). Springer International Publishing.

[Fuchs, 2005] Fuchs, J. J. (2005). Recovery of exact sparse representations in the presence of bounded noise. *Information Theory, IEEE Transactions on*, 51(10), 3601-3608.

[Gan et al., 2011] Gan, L., Tu, W., Liu, G., & Yi, Y. (2011, March). Integrating Cliques as Query Context into Information Retrieval Model. In *Proceedings of the 2011 Third International Workshop on Education Technology and Computer Science*, Volume 01 (pp. 110-113). IEEE Computer Society.

[Gavankar et al., 2012] Gavankar, C., Kulkarni, A., Fang Li, Y. and Ramakrishnan, G. (2012), Enriching an Academic Knowledge base using Linked Open Data. *Proceedings of the Workshop on Speech and Language Processing Tools in Education in 24th International Conference on Computational Linguistics*, pp. 51-60.

[van Gemert et al., 2008] van Gemert, J. C., Geusebroek, J. M., Veenman, C. J., & Smeulders, A. W. (2008). Kernel codebooks for scene categorization. *Computer Vision–ECCV 2008* (pp. 696-709). Springer Berlin Heidelberg.

[van Gemert et al., 2010] van Gemert, J. C., Veenman, C. J., Smeulders, A. W., & Geusebroek, J. M. (2010). Visual word ambiguity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7), 1271-1283.

[Ghawi & Cullot, 2007] Ghawi, R. and Cullot, N. (2007), Database-to-Ontology Mapping Generation for Semantic Interoperability. *VLDB '07 Conference, VLDB Endowment ACM*, pp. 1-8, Vienna, Austria.

[Gini, 1921] Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, 124-126.

[Gisin, 2009] Gisin, N. (2009). Bell Inequalities: Many Questions, a Few Answers. *Quantum Reality, Relativistic Causality, and Closing the Epistemic Circle*. 73, 125-138.

[Goldman et al., 2013] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B., & Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* 494, 77–80.

[Golge & Duygulu, 2014] Golge, E., & Duygulu, P. (2014). ConceptMap: Mining Noisy Web Data for Concept Learning. *Computer Vision–ECCV 2014* (pp. 439-455). Springer International Publishing.

[Gómez-Pérez & Manzano-Macho, 2003] Gómez-Pérez, A. and Manzano-Macho, D. (2003), A survey of ontology learning methods and techniques. Onto-web IST Project, Deliverable 1.5. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.3714&rep=rep1&type=pdf.

[Grace & Knight, 2008] Grace, S. & Knight, G. (2008) What are significant properties and why should I care? *Presentation delivered at Digital Curation 101*, October, 7 2008. Edinburgh, Scotland.

[Grass et al., 2015] Grass, R.N., Heckel, R., Puddu, M., Paunescu, D., & Stark, W.J. (2015). Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angewandte Chemie International Edition*, 54, 2552–2555. Available at: http://goo.gl/GVfiLD.

[Gray & Neuhoff, 1998] Gray, R. M., & Neuhoff, D. L. (1998). Quantization. *Information Theory, IEEE Transactions on*, 44(6), 2325-2383.

[Han et al., 2008] Han, L., Finin, T., Parr, C., Sachs, J. and Joshi, A. (2008), RDF123: from Spreadsheets to RDF. *Proceedings of the 7th International Semantic Web Conference*, pp. 451-466, Springer Berlin Heidelberg.

[Hariharan et al., 2014] Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2014). Simultaneous detection and segmentation. *Computer Vision–ECCV 2014* (pp. 297-312). Springer International Publishing.

[Harris, 1954] Harris, Z. 1954. Distributional structure. *Word* 10 (23), 146-162.

[Harris & Stephens, 1998] Harris, C., & Stephens, M. (1988, August). A combined corner and edge detector. *Alvey Vision Conference* (Vol. 15, p. 50).

[He et al., 2004] He, X., Zemel, R. S., & Carreira-Perpiñán, M. Á. (2004, June). Multiscale conditional random fields for image labeling. *Proc. 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2004)*, Vol. 2, pp. II-695. IEEE.

[Hoi & Lyu, 2005] Hoi, S. C., & Lyu, M. R. (2005, June). A semi-supervised active learning framework for image retrieval. *Proc. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2005)*, Vol. 2, pp. 302-309. IEEE.

[Hosmer & Lemeshow, 2004] Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

[Hoyer, 2004] Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5, 1457-1469.

[Huang & Aviyente, 2006] Huang, K., & Aviyente, S. (2006). Sparse representation for signal classification. *Advances in neural information processing systems* (pp. 609-616).

[Huang et al., 2013] Huang, J., Liu, H., Shen, J., & Yan, S. (2013). Towards efficient sparse coding for scalable image annotation. *Proc. of 21st ACM Int. Conf. on Multimedia*, pp. 947-956. ACM.

[Huiskes et al., 2010] Huiskes, M. J., Thomee, B., & Lew, M. S. (2010, March). New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative. *Proc. Int. Conf. on Multimedia information retrieval*, pp. 527-536. ACM.

[Hurley & Rickard, 2009] Hurley, N., & Rickard, S. (2009). Comparing measures of sparsity. *Information Theory, IEEE Transactions on*, 55(10), 4723-4741.

[Jardine & van Rijsbergen, 1971] Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5), 217-240.

[Jégou et al., 2010] Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010, June). Aggregating local descriptors into a compact image representation. *Proc. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3304-3311. IEEE.

[Jégou et al., 2011] Jégou, H., Douze, M., & Schmid, C. (2011). Product quantization for nearest neighbor search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1), 117-128.

[Jeong, 2013] Jeong, D. (2013), Implementation of Ontology Learning and Population System from Structured Data Sources: Standard-based Approach. *International Journal of Software Engineering and its Applications*, Vol. 7, No. 6, pp. 289-304.

[Jiang & Conrath, 1997] Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

[Jiménez et al., 2009] Jiménez, V., Vilanova, L., Gelado, I., Gil, M., Fursin, G., and Navarro, N. (2009). Predictive runtime code scheduling for heterogeneous architectures. *High Performance Embedded Architectures and Compilers*, pages 19–33.

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features (pp. 137-142). Springer Berlin Heidelberg.

[Johnson, 1967] Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3), 241-254.

[Jolliffe, 2002] Jolliffe, I. (2002). *Principal component analysis*. John Wiley & Sons, Ltd.

[Jovanović et al., 2007] Jovanović, J., Gašević, D., Knight, C., & Richards, G. (2007). Ontologies for effective use of context in e-learning settings. *Journal of Educational Technology & Society*, 10(3), 47-59.

[Karvanen & Cichocki, 2003] Karvanen, J., & Cichocki, A. (2003, April). Measuring sparseness of noisy signals. *Proc. 4th Int. Symposium on Independent Component Analysis and Blind Signal Separation* (pp. 125-130).

[Khrennikov, 2010] Khrennikov, A. (2010). *Ubiquitous Quantum Structure: From Psychology to Finance*. Springer: New York.

[Kirk & Hwu, 2009] Kirk, D. and Hwu, W. (2009). *Programming massively parallel processors: A hands-on approach*.

[Kohonen, 2001] Kohonen, T. (2001). *Self-Organizing Maps*. Springer.

[Koop et al., 2006] Koop, M., Sur, S., Gao, Q., and Panda, D. (2006). High performance MPI design using unreliable datagram for ultra-scale InfiniBand clusters. *Proc. 21st Annual Int. Conf. on Supercomputing (ISC-06)*, 180–189, Dresden, Germany.

[Kordumova et al., 2014] Kordumova, S., Li, X., & Snoek, C. G. (2014). Best practices for learning video concept detectors from social media examples. *Multimedia Tools and Applications*, 74(4), 1291-1315.

[Kreutz-Delgado & Rao, 1998] Kreutz-Delgado, K., & Rao, B. D. (1998, May). Measures and algorithms for best basis selection. *Proc. 1998 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 3, pp. 1881-1884. IEEE.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (pp. 1097-1105).

[Kuhn et al., 2000] Kuhn, B., Petersen, P., and O'Toole, E. (2000). OpenMP versus threading in C/C++. *Concurrency: Practice and Experience*, 12(12):1165–1176.

[Lahabar & Narayanan, 2009] Lahabar, S. and Narayanan, P. (2009). Singular value decomposition on GPU using CUDA. *Proc. 23rd Int. Symposium on Parallel and Distributed Processing (IPDPS-09)*, Rome, Italy.

[Landauer & Dumais, 1997] Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

[Lau et al., 2008] Lau, R. Y., Bruza, P. D., & Song, D. (2008). Towards a belief-revision-based adaptive and context-sensitive information retrieval system. *ACM Transactions on Information Systems (TOIS)*, 26(2), 8.

[Lazebnik et al., 2006] Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proc. 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Vol. 2, pp. 2169-2178. IEEE.*

[Lee et al., 2009] Lee, S., Min, S., and Eigenmann, R. (2009). OpenMP to GPGPU: a compiler framework for automatic translation and optimization. *Proc. 14th Symposium on Principles and Practice of Parallel Programming (PPOPP-09), pp. 101–110.*

[Lerman et al., 2011] Lerman, K., Hogg, T., Galstyan, A., and Steeg, G. V. (2011). Social Mechanics: An Empirically Grounded Science of Social Media. *Proc. ICWSM workshop on the Future of Social Media (FOSW11).*

[Lewis, 1999] Lewis, D. (1999). Reuters-21578 text categorization test collection distribution 1.0.

[Li et al., 2010] Li, Q., Kecman, V. & Salman, R. (2010). A Chunking Method for Euclidean Distance Matrix Calculation on Large Dataset Using Multi-GPU. *Proceedings of ICMLA-10, 9th International Conference on Machine Learning and Applications*, 208-213.

[Li et al., 2013] Li, X., Snoek, C. G., Worring, M., Koelma, D., & Smeulders, A. W. (2013). Bootstrapping visual categorization with relevant negatives. *Multimedia, IEEE Transactions on*, 15(4), 933-945.

Li, Q.; Kecman, V. & Salman, R. A Chunking Method for Euclidean Distance Matrix Calculation on Large Dataset Using Multi-GPU. Proceedings of ICMLA-10, 9th International Conference on Machine Learning and Applications, 2010, 208-213.

[Li & Guo, 2013] Li, X., & Guo, Y. (2013, June). Adaptive active learning for image classification. *Proc. 2013 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 859-866. IEEE.

[Li et al., 2014] Li, W., Niu, L., & Xu, D. (2014). Exploiting privileged information from web data for image categorization. *Computer Vision–ECCV 2014* (pp. 437-452). Springer International Publishing.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91-110.

[Luk et al., 2009] Luk, C., Hong, S., and Kim, H. (2009). Qilin: Exploiting parallelism on heterogeneous multiprocessors with adaptive mapping. *Proc. 42nd Annual IEEE/ACM Int. Symposium on Microarchitecture (MICRO-42)*, pages 45–55, New York, NY, USA.

[Lund et al., 1995] Lund, K., Burgess, C., & Atchley, R.A. (1995). Semantic and associative priming in high-dimensional semantic space. *Proc. 17th Annual Conf. of the Cognitive Science Society*, pages 660-665.

[Luo et al., 2005] Luo, Z., Liu, H., Yang, Z., and Wu, X. (2005). Self-organizing maps computing on graphic process unit. Proc. *13th European Symposium on Artificial Neural Networks (ESANN-05)*.

[Maamar et al., 2006] Maamar, Z., Benslimane, D., & Narendra, N. C. (2006). What can context do for web services?. *Communications of the ACM*, 49(12), 98-103.

[Maedche & Staab, 2004] Maedche, A., & Staab, S. (2004). Ontology learning. *Handbook on ontologies* (pp. 173-190). Springer Berlin Heidelberg.

[Mairal et al., 2007] Mairal, J., Sapiro, G., & Elad, M. (2007). *Learning multiscale sparse representations for image and video restoration* (No. IMA-PREPRINT-SERIES-2168). MINNESOTA UNIV MINNEAPOLIS INST FOR MATHEMATICS AND ITS APPLICATIONS.

[Mairal et al., 2008] Mairal, J., Elad, M., & Sapiro, G. (2008). Sparse representation for color image restoration. *Image Processing, IEEE Transactions on*, 17(1), 53-69.

[Manjunath et al., 2001] Manjunath, B. S., Ohm, J. R., Vinod, V. V., and Yamada, A. (2001) "Colour and texture descriptors," *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7*, vol. 11, no. 6, pp. 703–715, Jun 2001.

[Maryak & Chin, 2001] Maryak, J. L., & Chin, D. C. (2001). Global random optimization by simultaneous perturbation stochastic approximation. *Proc. 2001 American Control Conference*, Vol. 2, pp. 756-762. IEEE.

[Masada et al., 2009] Masada, T., Hamada, T., Shibata, Y., and Oguri, K. (2009). Accelerating collapsed variational Bayesian inference for latent Dirichlet allocation with NVidia CUDA compatible devices. *Proc. IEA/AIE-09, 22nd Int. Conf. on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 491–500, Tainan, Taiwan.

[Matas et al., 2004] Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10), 761-767.

[Mayer, 2004] Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning?. *American Psychologist*, 59(1), 14.

[Maynard et al., 2009] Maynard, D., Funk, A. and Peters, W. (2009), SPRAT: A tool for automatic semantic pattern-based ontology population. *International Conference for Digital Libraries and the Semantic Web*, Trento, Italy.

[McAuley & Leskovec, 2013] McAuley, J. and Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. *Proc. 7th ACM Conference on Recommender Systems (RecSys-13)*, pp. 165–172. ACM.

[Meinshausen & Yu, 2009] Meinshausen, N., & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 246-270.

[Mikolajczyk & Schmid, 2005] Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10), 1615-1630.

[Modica et al., 2001] Modica, G., Gal, A. and Jamil, H.M. (2001), The Use of Machine-Generated Ontologies in Dynamic Information Seeking. *Cooperative Information Systems*, pp. 433-447, Springer Berlin Heidelberg.

[Morstatter et al., 2013] Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. arXiv preprint arXiv:1306.5204.

[Moshfeghi, 2012] Moshfeghi, Y. (2012). Role of emotion in information retrieval (Doctoral dissertation, University of Glasgow).

[Mrissa et al., 2009] Mrissa, M., Dietze, S., Thiran, P., Ghedira, C., Benslimane, D., & Maamar, Z. (2009). Context-based semantic mediation in web service communities. *Weaving Services and People on the World Wide Web* (pp. 49-66). Springer Berlin Heidelberg.

[Ng & Cardie, 2003] Ng, V., & Cardie, C. (2003, July). Bootstrapping coreference classifiers with multiple machine learning algorithms. *Proc. 2003 Conf. on Empirical methods in natural language processing*, pp. 113-120. Association for Computational Linguistics.

[NISO, 2010] National Information Standards Organization (NISO). (Spring 2010). *Information Standards Quarterly (ISQ) - Special Issue: Digital Preservation*. Volume 22, Issue 2.

[Nowak & Rüger, 2010] Nowak, S., & Rüger, S. (2010, March). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. *Proc. Int. Conf. on Multimedia Information Retrieval*, pp. 557-566. ACM.

[NVidia, 2014] NVidia Corporation. NVidia Compute Unified Device Architecture Programming Guide 6.0, 2014.

[O'Donnell et al., 2001] O'Donnell, M., Mellish, C., Oberlander, J., & Knott, A. (2001). ILEX: an architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7(03), 225-250.

[Olshausen & Field, 2004] Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4), 481-487.

[Oquab et al., 2014a] Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014, June). Learning and transferring mid-level image representations using convolutional neural networks. *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* (pp. 1717-1724). IEEE.

[Oquab et al., 2014b] Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Weakly supervised object recognition with convolutional neural networks.

[Owens et al., 2007] Owens, J., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A., and Purcell, T. (2007). A survey of general-purpose computation on graphics hardware. *Computer Graphics Forum*, 26(1):80–113.

[Pang, B., & Lee, 2008] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2), 1-135.

[Papadopoulou & Mezaris, 2015] Papadopoulou, O., & Mezaris, V. (2015, June). Exploiting Multiple Web Resources towards Collecting Positive Training Samples for Visual Concept Learning. *Proceedings of the 5th ACM on Int. Conf. on Multimedia Retrieval* (pp. 531-534). ACM.

[Pastor et al., 2013] Pastor, G., Mora-Jiménez, I., Jäntti, R., & Caamano, A. J. (2013, August). Sparsity-Based criteria for entropy measures. *Wireless Communication Systems (ISWCS 2013), Proceedings of the Tenth International Symposium on* (pp. 1-5). VDE.

[PERICLES D2.3.1, 2014] PERICLES Consortium, Deliverable 2.3.1: Media and science case study functional requirements and user descriptions, June 2014.

[PERICLES D2.3.2, 2015] PERICLES Consortium, Deliverable 2.3.2: Data Survey and Domain Ontologies for Case Studies, September 2015.

[PERICLES D3.3, 2015] PERICLES Consortium, Deliverable 3.3: Semantics for Change Management, July 2015.

[PERICLES D4.1, 2014] PERICLES Consortium, Deliverable 4.1: Initial version of Environment Information Extraction Tools, July 2014.

[PERICLES D4.2, 2015] PERICLES Consortium, Deliverable 4.2: Encapsulation of environmental information, November 2015.

[PERICLES D5.2, 2015] PERICLES Consortium, Deliverable 5.2: Basic Tools for Digital Ecosystem Management, November 2015.

[Perronnin et al., 2010] Perronnin, F., Sánchez, J., & Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. *Computer Vision–ECCV 2010* (pp. 143-156). Springer Berlin Heidelberg.

[Petatsis et al., 2011] Petatsis, G., Karkaletsis, V., Paliouras, G., Krithara, A., & Zavitsanos, E. (2011, January). Ontology population and enrichment: State of the art. *Knowledge-driven multimedia information extraction and ontology evolution* (pp. 134-166). Springer-Verlag.

[Phillips, 1997] Phillips, D. C. (1997). How, why, what, when, and where: Perspectives on constructivism in psychology and education. *Issues in Education*,3(2), 151-194.

[Phillips et al., 2008] Phillips, J., Stone, J., and Schulten, K. (2008). Adapting a message-driven parallel application to GPU-accelerated clusters. *Proc. 21st Conf. on Supercomputing (SC-08)*, pp. 1-9, Austin, TX, USA.

[Piaget, 1970] Piaget, J. (1970). Science of education and the psychology of the child. Trans. D. Coltman.

[Platt, 1999] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3), 61-74.

[Priem et al., 2010] Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto.

[Rao & Kreutz-Delgado, 1999] Rao, B. D., & Kreutz-Delgado, K. (1999). An affine scaling methodology for best basis selection. *Signal Processing, IEEE Transactions on*, *47*(1), 187-200.

[Rath et al., 2008] Rath, G., Guillemot, C., & Fuchs, J. J. (2008, October). Sparse approximations for joint source-channel coding. *Multimedia Signal Processing, 2008 IEEE 10th Workshop on* (pp. 481-485). IEEE.

[Ritter et al., 2011] Ritter, A., Clark, S., & Etzioni, O. (2011, July). Named entity recognition in tweets: an experimental study. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1524-1534). Association for Computational Linguistics.

[Rodrigues et al., 2014] Rodrigues, F., Pereira, F., & Ribeiro, B. (2014). Gaussian Process Classification and Active Learning with Multiple Annotators. *Proceedings of the 31st International Conference on Machine Learning* (ICML-14) (pp. 433-441).

[van de Sande et al., 2011] van de Sande, K., Gevers, T., and Snoek, C. (2011). Empowering visual categorization with the GPU. *IEEE Transactions on Multimedia*, 13(1):60–70.

[Sadegh & Spall, 1998] Sadegh, P., & Spall, J. C. (1998). Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation. *Automatic Control, IEEE Transactions on*, 43(10), 1480-1484.

[van de Sande et al., 2010] Van De Sande, K. E., Gevers, T., & Snoek, C. G. (2010). Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9), 1582-1596.

[Sanderson & Watry, 2007] Sanderson, R. and Watry, P. (2007). Integrating data and text mining processes for digital library applications. *Proceedings of JCDL-07, 7th ACM/IEEE- CS Joint Conference on Digital Libraries*, pages 73–79, Vancouver, Canada.

[Sahlgren, 2006] Sahlgren, M. (2006). The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.

[Sánchez & Perronnin, 2011] Sánchez, J., & Perronnin, F. (2011, June). High-dimensional signature compression for large-scale image classification. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 1665-1672). IEEE.

[Schilit & Theimer, 1994] Schilit, B. N., & Theimer, M. M. (1994). Disseminating active map information to mobile hosts. *Network*, IEEE, 8(5), 22-32.

[Schlieder, 2010] Schlieder, C. 2010. Digital heritage: Semantic challenges of long-term preservation. *Semantic Web* (1)1-2, 143-147.

[Schmidt et al., 1999] Schmidt, A., Beigl, M., & Gellersen, H. W. (1999). There is more to context than location. *Computers & Graphics*, *23*(6), 893-901.

[Settles, 2010] Settles, B. (2010). Active learning literature survey. University of Wisconsin, Madison, 52(55-66), 11.

[Shi et al., 2009] Shi, Q., Petterson, J., Dror, G., Langford, J., Smola, A., & Vishwanathan, S. V. N. (2009). Hash kernels for structured data. *The Journal of Machine Learning Research*, 10, 2615-2637.

[Shamma, 2014] Shamma, D. A. (2014). One hundred million creative commons Flickr images for research.

[Spall, 1999] Spall, J. C. (1999, December). Stochastic optimization and the simultaneous perturbation method. *Proceedings of the 31st conference on Winter simulation: Simulation---a bridge to the future*-Volume 1 (pp. 101-109). ACM.

[Stone et al., 2010] Stone, J., Hardy, D., Ufimtsev, I., and Schulten, K. (2010). GPU-accelerated molecular modeling coming of age. *Journal of Molecular Graphics and Modelling*, 29(2):116–125.

[Strang et al., 2003] Strang, T., Linnhoff-Popien, C., & Frank, K. (2003, January). CoOL: A context ontology language to enable contextual interoperability. *Distributed applications and interoperable systems* (pp. 236-247). Springer Berlin Heidelberg.

[Strong & Gong, 2008] Strong, G. and Gong, M. (2008). Browsing a large collection of community photos based on similarity on GPU. *Advances in Visual Computing*, pages 390–399.

[Sul & Tovchigrechko, 2011] Sul, S. and Tovchigrechko, A. (2011). Parallelizing BLAST and SOM algorithms with MapReduce-MPI library. *Proceedings of IPDPS-11, 25th International Parallel and Distributed Computing Symposium*, pages 476–483, Anchorage, AK, USA.

[Sun et al., 2013] Sun, X., Wang, H. and Yu, Y. (2013), SAPOP: Semiautomatic Framework for Practical Ontology Population from Structured Knowledge Bases. *Semantic Web and Web Science*, pp. 181-186, Springer New York.

[Tenenbaum et al., 2000] Tenenbaum, J., Silva, V., and Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

[Thelwall et al., 2013] Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services.

[Thomee & Popescu, 2012] Thomee, B., & Popescu, A. (2012). Overview of the clef 2012 flickr photo annotation and retrieval task. *In the working notes for the CLEF 2012 labs/workshop*. Rome, Italy.

[Tipping, 2001] Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *The journal of machine learning research*, 1, 211-244.

[Trier, 1934] Trier, J. (1934). Das sprachliche Feld. *Neue Jahrbücher fur Wissenschaft und Jugendbildung*, 10:428–449.

[Tsirelson, 1980] Tsirelson, B. (1980). Quantum Generalizations of Bell's Inequality. *Letters in Mathematical Physics*, 4, 93-100.

[Turney, 2002] Turney, P.D. (2002), Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, Pennsylvania, pp. 417-424.

[Turney & Pantel, 2010] Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

[Ufimtsev & Martınez, 2008] Ufimtsev, I. and Martinez, T. (2008). Graphical processing units for quantum chemistry. *Computing in Science & Engineering*, 10(6):26–34.

[Ultsch & Mörchen, 2005] Ultsch, A. and Mörchen, F. (2005). ESOM-Maps: Tools for clustering, visualization, and classification with emergent SOM. *Technical report*, Data Bionics Research Group, University of Marburg.

[Vedaldi et al., 2007] Vedaldi, A., Fulkerson, B., & Feat, V. L. (2007). An open and portable library of computer vision algorithms. *Proceedings of the 18th annual ACM international conference on Multimedia* (pp. 1469-1472).

[Velardi et al., 2002] Velardi, P., Navigli, R. and Missikoff, M. (2002), Integrated approach for Web ontology learning and engineering. *IEEE Computer*, Vol. 35, No. 11, pp. 60-63.

[Vijayanarasimhan & Grauman, 2014] Vijayanarasimhan, S., & Grauman, K. (2014). Large-scale live active learning: Training object detectors with crawled data and crowds. *International Journal of Computer Vision*, 108(1-2), 97-114.

[Vygotsky, 1980] Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard University Press.

[Wang et al., 2004] Wang, X. H., Zhang, D. Q., Gu, T., & Pung, H. K. (2004, March). Ontology based context modeling and reasoning using OWL. *Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on* (pp. 18-22). IEEE.

[Wang et al., 2009] Wang, C., Yan, S., Zhang, L., & Zhang, H. J. (2009, June). Multi-label sparse coding for automatic image annotation. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 1643-1650). IEEE.

[Wang et al., 2010] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. (2010, June). Locality-constrained linear coding for image classification. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (pp. 3360-3367). IEEE.

[Wehrens & Buydens, 2007] Wehrens, R. and Buydens, L. M. C. (2007). Self- and super-organizing maps in R: The kohonen package. *Journal of Statistical Software*, 21(5):1–19.

[Wei & JaJa, 2011] Wei, Z. and JaJa, J. (2011). A fast algorithm for constructing inverted files on heterogeneous platforms. *Proceedings of IPDPS-11, 25th International Parallel and Distributed Computing Symposium*, Anchorage, AK, USA.

[Weinberger et al., 2004] Weinberger, K., Sha, F., and Saul, L. (2004). Learning a kernel matrix for nonlinear dimensionality reduction. *Proceedings of ICML-04, 21st International Conference on Machine learning*, pages 106–113.

[Weinberger et al., 2009] Weinberger, K., Dasgupta, A., Langford, J., Smola, A., & Attenberg, J. (2009, June). Feature hashing for large scale multitask learning. *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1113-1120). ACM.

[Weiser, 1993] Weiser, M. (1993). Some computer science issues in ubiquitous computing. *Communications of the ACM*, *36*(7), 75-84.

[Wittgenstein, 1953] Wittgenstein, L. 1953. *Philosophical investigations*. Oxford: Blackwell.

[Wittek, 2014] Wittek, P. (2014). Quantum machine learning. Elsevier: Amsterdam.

[Wittek & Darányi, 2012] Wittek, P. and Darányi, S. (2012). A GPU-accelerated algorithm for self- organizing maps in a distributed environment. *Proceedings of ESANN-12, 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

[Wittek et al., 2014] Wittek, P., Darányi, S., and Lin, Y. (2014). A vector field approach to lexical semantics. *Proceedings of QI-14, 8th International Conference on Quantum Interaction*, pages 78–92.

[Wittek, 2015] Wittek, P. (2015). Algorithm 950: Ncpol2sdpa---Sparse Semidefinite Programming Relaxations for Polynomial Optimization Problems of Noncommuting Variables. *ACM Transactions on Mathematical Software*, 41, 21.

[Wittek et al., 2015a] Wittek, P., Darányi, S. Kontopoulos, E., Moysiadis, T., Kompatsiaris, I. 2015. *Monitoring Term Drift Based on Semantic Consistency in an Evolving Vector Field*. Available at http://arxiv.org/abs/1502.01753.

[Wittek et al., 2015b] Wittek, P., Gao, S. C., Lim, I. S., and Zhao, L. (2015). Somoclu: An efficient parallel library for self-organizing maps. arXiv:1305.1422.

[Wright et al., 2009] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(2), 210-227.

[Xi et al., 2009] Xi, Y. T., Xiang, Z. J., Ramadge, P. J., & Schapire, R. E. (2009). Speed and sparsity of regularized boosting. International Conference on Artificial Intelligence and Statistics (pp. 615-622).

[Yan et al., 2009] Yan, F., Xu, N., and Qi, Y. (2009). Parallel inference for latent Dirichlet allocation on graphics processing units. *Advances in Neural Information Processing Systems*, 22.

[Yan et al., 2011] Yan, Y., Fung, G. M., Rosales, R., & Dy, J. G. (2011). Active learning from crowds. *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 1161-1168).

[Yang et al., 2010] Yang, J., Bouzerdoum, A., Tivive, F. R. C., & Phung, S. L. (2010, July). Dimensionality reduction using compressed sensing and its application to a large-scale visual recognition task. *Neural Networks (IJCNN), The 2010 International Joint Conference on* (pp. 1-8). IEEE.

[Yu et al., 2012] Yu, S., Tranchevent, L. C., Liu, X., Glänzel, W., Suykens, J. A., De Moor, B., & Moreau, Y. (2012). Optimized data fusion for kernel k-means clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5), 1031-1039.

[Zhang & Li, 2010] Zhang, Q., & Li, B. (2010, June). Discriminative K-SVD for dictionary learning in face recognition. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (pp. 2691-2698). IEEE.

[Zhang et al., 2011] Zhang, L., Ma, J., Cui, C., & Li, P. (2011, April). Active learning through notes data in flickr: an effortless training data acquisition approach for object localization. *Proceedings of the 1st ACM International Conference on Multimedia Retrieval* (p. 46). ACM.

[Zhang et al., 2012] Zhang, Z., Wang, J., & Zha, H. (2012). Adaptive manifold learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(2), 253-265.

[Zhao & Ichise, 2012] Zhao, L. and Ichise, R. (2012), Mid-Ontology Learning from Linked Data. *The Semantic Web*, pp. 112-127, Springer Berlin Heidelberg.

[Zhou et al., 2010] Zhou, X., Yu, K., Zhang, T., & Huang, T. S. (2010). Image classification using super-vector coding of local image descriptors. *Computer Vision–ECCV 2010* (pp. 141-154). Springer Berlin Heidelberg.

[Zonoobi et al., 2011] Zonoobi, D., Kassim, A., & Venkatesh, Y. V. (2011). Gini index as sparsity measure for signal reconstruction from compressive samples. *Selected Topics in Signal Processing, IEEE Journal of*, 5(5), 927-932.

[Zubiaga et al., 2011] Zubiaga, A., Spina, D., Fresno, V., & Martínez, R. (2011, October). Classifying trending topics: a typology of conversation triggers on twitter. *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 2461-2464). ACM.

Appendix

PROOF OF THE THREE THEOREMS MENTIONED IN THIS DELIVERABLE

Theorem 1

$0 \leq S_p(\mathbf{c}) \leq 1 - \frac{1}{N}, \ \forall \mathbf{c} \in \mathbb{R}^N, \ \forall p \geq 1$

Proof.

$S_p(\mathbf{c}) = \alpha \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (c_j - c_i)^p ,$

where

$\alpha = \frac{1}{N \sum_{i=1}^{N} c_i^p}$

But

$(c_j - c_i)^p \leq c_j^p - c_i^p$

$c_j \geq c_i \geq 0$

Recall that the coefficients are sorted. Therefore

$S_p(\mathbf{c}) \leq \alpha \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (c_j^p - c_i^p)$

$= \alpha \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} c_j^p - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} c_i^p \right]$

$= \alpha \Big[(c_2^p + c_3^p + \cdots + c_N^p) + (c_3^p + \cdots + c_N^p) + \cdots + c_N^p$

$-(N-1)c_1^p - (N-2)c_2^p - \cdots - 2c_{N-2}^p - c_{N-1}^p \Big]$

$\leq \alpha \Big[(c_2^p + c_3^p + \cdots + c_N^p) + (c_3^p + \cdots + c_N^p) + \cdots + c_N^p \Big]$

$= \alpha \Big[c_2^p + 2c_3^p + \cdots + (N-1)c_N^p \Big]$

$\leq \alpha \Big[(N-1)c_1^p + (N-1)c_2^p + \cdots + (N-1)c_N^p \Big]$

$= \alpha (N-1) \sum_{i=1}^{N} c_i^p = \frac{N-1}{N} = 1 - \frac{1}{N}.$

Theorem 2

Given a vector

$\mathbf{c} = [c_1, c_2, \ldots, c_N]$

if $p>q$, then

$S_p(\mathbf{c}) < S_q(\mathbf{c})$

Proof.

For proving this theorm we can assume that

$c_i > 1, \forall i \in \{1, 2, \ldots, N\}$

Otherwise, since from Theorem 5 we have that

$S_p(\mathbf{c}) = S_p(\alpha \mathbf{c}), \alpha > 0$

we can multiply each coefficient c_i by a value $\alpha > 1/c_1$ to obtain

$\alpha c_i > 1, \forall i \in \{1, 2, \ldots, N\}.$

We have that

$$\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(c_j-c_i)^p}{c_N^p} = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{(c_j-c_i)^p}{c_N^p}$$
$$= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\frac{c_j-c_i}{c_N}\right)^p < \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\frac{c_j-c_i}{c_N}\right)^q$$
$$= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{(c_j-c_i)^q}{c_N^q} = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(c_j-c_i)^q}{c_N^q},$$

where the inequality holds because

$\frac{c_j-c_i}{c_N} < 1$

and *p>q*

Moreover, since we assumed that

$c_i > 1, \forall i \in \{1, 2, \ldots, N\},$

we also have that

$c_{N-1}^p + c_{N-2}^p + \cdots + c_1^p > c_{N-1}^q + c_{N-2}^q + \cdots + c_1^q,$

which combined with inequality .. straightforwardly implies that

$\frac{1}{N}\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(c_j-c_i)^p}{c_N^p+(c_{N-1}^p+\cdots+c_1^p)} < \frac{1}{N}\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(c_j-c_i)^q}{c_N^q+(c_{N-1}^q+\cdots+c_1^q)},$

which is equivalent to

$S_p(\mathbf{c}) < S_q(\mathbf{c}).$

Theorem 3

For an arbitrary vector **c**, we have that

$\lim_{p\to+\infty} S_p(\mathbf{c}) = 0$

Proof.

For every *p* we have that

$$0 \le S_p(\mathbf{c}) = \frac{1}{N}\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(c_j - c_i)^p}{\sum_{i=1}^{N} c_i^p}$$
$$\le \frac{1}{N}\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(c_j - c_i)^p}{c_N^p} = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{(c_j - c_i)^p}{c_N^p}$$
$$= \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\frac{c_j - c_i}{c_N}\right)^p.$$

But since

$c_j - c_i < c_N, \forall i, j$

we have that

$$\lim_{p\to+\infty}\left(\frac{c_j - c_i}{c_N}\right)^p = 0, \forall i, j,$$

which implies that

$$\lim_{p\to+\infty}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\frac{c_j - c_i}{c_N}\right)^p = 0.$$

Hence, applying the squeeze theorem in inequality,

$$\lim_{p\to+\infty} S_p(\mathbf{c}) = 0$$

PROOF THAT GDS OF FIRST ORDER IS EQUIVALENT TO GI

Theorem 4

GDS of order 1 is equivalent to GI.

Proof.

$$S_1(\mathbf{c}) = \frac{1}{N\|\mathbf{c}\|_1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(c_j - c_i)$$
$$= \frac{1}{N\|\mathbf{c}\|_1}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_j - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_i\right)$$
$$= \frac{1}{N\|\mathbf{c}\|_1}\left(\sum_{i=1}^{N}(i-1)c_i - \sum_{i=1}^{N}(N-i)c_i\right)$$
$$= \frac{1}{N\|\mathbf{c}\|_1}\left(-N\sum_{i=1}^{N} c_i - \sum_{i=1}^{N} c_i + 2\sum_{i=1}^{N} ic_i\right)$$
$$= \frac{1}{\|\mathbf{c}\|_1}\left(\sum_{i=1}^{N} c_i - 2\sum_{i=1}^{N} c_i + \frac{2}{N}\sum_{i=1}^{N} ic_i - \frac{1}{N}\sum_{i=1}^{N} c_i\right)$$
$$= 1 - \frac{2}{\|\mathbf{c}\|_1}\left(\sum_{i=1}^{N} c_i - \frac{1}{N}\sum_{i=1}^{N} ic_i + \frac{1}{2N}\sum_{i=1}^{N} c_i\right)$$
$$1 - \frac{2}{\|\mathbf{c}\|_1}\sum_{i=1}^{N} c_i\left(1 - \frac{i}{N} + \frac{1}{2N}\right) = GI(\mathbf{c})$$

SPARSITY METRIC OBJECTIVE CRITERIA

- **P1: Continuity:**

S(c + dc) → S(c), when *dc → 0*.

This property requires that small changes of the coefficients should not lead to dramatic change of sparsity.

- **P2: Permutation Invariance:**

*S(t(**c**)) = S(**c**)*, where *t(**c**)* is a permutation of **c**.

This property is very important and actually postulates that permuting the coefficients of a signal should not affect its sparsity. Provided that this property holds for a metric – which is usually the case – for convenience and without loss of generality, given an arbitrary vector, since the position of the coefficients does not matter, we can consider that these have been sorted in ascending order, i.e.: $0 \le c_1 \le c_2 \le \cdots \le c_N$.

- **P3: Robin Hood:**

Let **c =** $[c_1,c_2,\ldots,c_N]$. If **c'** = $[c_1,\ldots,c_i + a,\ldots,c_j - a,\ldots,c_N]$, then *S(**c'**) < S(**c**)*, for all a,c_i,c_j, such that $c_j > c_i$ and $0<a<(c_j-c_i)/2$.

Robin Hood property says that subtracting a specific amount from a large coefficient and adding this amount to a smaller coefficient decreases the vector-sparsity as the energy of the signal spreads out along the coefficients. The constraint of α is used to avoid $c_i > c_j$.

- **P4: Scaling:**

*S(a**c**) = S(**c**)*, $\forall a \in R, a > 0$.

Scaling property requires that by multiplying all vector coefficients with the same scalar must not affect vector-sparsity.

- **P5: Rising Tide:**

*S(a + **c**) < S(**c**)*, $a \in R, a > 0$, except for the case where $c_1 = c_2 = \cdots = c_N$.

Rising Tide says that adding the same scalar to all vector coefficients reduces vector-sparsity. The significance of this property becomes more obvious by examining the limit behaviour of the signal under this operation. Indeed, by adding an increasing amount to all coefficients, the relative difference among coefficients becomes negligible and therefore the sparsity should asymptotically become zero.

- **P6: Cloning:**

*S(***c***) = S(***c***|***c***) = ⋯ = S(***c***|***c***|…|***c***)*, where | denotes concatenation.
Cloning requires that concatenating a number of vectors, which comprise exact copies of the original one must not affect vector-sparsity. This is also quite reasonable, if we take again into account that the relative difference of the signal coefficients after sorting the resulting signal is kept intact.

- **P7: Bill Gates:**

$\forall i \in \{1,2,\ldots,N\}$, and $\forall a > 0 : S([c_1,\ldots,c_i + a,\ldots,c_n]) > S([c_1,\ldots,c_i,\ldots,c_n])$.
By increasing the value of a vector coefficient, while maintaining the remaining coefficients, the sparsity increases, as the vector energy is concentrated to a mere coefficient.

- **P8: Babies:**

*S(***c***|0) > S(***c***)*.
By adding extra zero's to the original vector, the vector sparsity increases. This action has similar effect to P7, since by adding zero's the energy is concentrated to fewer coefficients.

- **P9: Saturation:**

$$\lim_{N\to+\infty} \frac{S(\mathbf{0}_N||1)}{S(\mathbf{0}_{N-1}||1)} = 1.$$

Saturation says that by concatenating extra zeros to a vector, the change of its sparsity asymptotically becomes negligible.

PROOF THAT GDS SATISFIES SPARSITY METRIC OBJECTIVE CRITERIA

In the following theorems, we rigorously prove that any GDS metric-instance satisfies all the aforementioned objective criteria for sparsity metrics. The ***Continuity*** and ***Permutation Invariance*** can be trivially proven from the definition. The proof of ***Robin Hood*** has been omitted due to its extensive length. Hereunder, we provide the proofs of properties P3 to P9.

Theorem 5

GDS satisfies **P4: Scaling**.

Proof.

$$S_p(a\mathbf{c}) = \frac{1}{N\left(\sum_{i=1}^N ac_i\right)^p} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (ac_j - ac_i)^p$$

$$= \frac{1}{a^p N\left(\sum_{i=1}^N c_i\right)^p} a^p \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (c_j - c_i)^p = S_p(\mathbf{c}).$$

Theorem 6

GDS satisfies **P5: Rising Tide**.

Proof.

First of all, if $c_1 = c_2 = \cdots = c_N$, then clearly $S_p(a + \mathbf{c}) = S_p(\mathbf{c}) = 0$. Otherwise, $S_p(\mathbf{c} + a) = \frac{1}{N\sum_{i=1}^N (c_i+a)^p} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} [(c_j + a) - (c_i + a)]^p =$

$\frac{1}{N\sum_{i=1}^N (c_i+a)^p} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (c_j - c_i)^p <$

$\frac{1}{N\sum_{i=1}^N c_i^p} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (c_j - c_i)^p = S_p(\mathbf{c})$, where the inequality follows from $\sum_{i=1}^N (c_i + a)^p > \sum_{i=1}^N c_i^p$, therefore $S_p(\mathbf{c} + a) < S_p(\mathbf{c})$.

Theorem 7

GDS satisfies **P6: Cloning**.

Proof.

$$S_p\left(\overbrace{\mathbf{c}\|\mathbf{c}\|\ldots\|\mathbf{c}}^{M-times}\right) =$$

$$S_p\left(\left[\overbrace{c_1,\ldots,c_1}^{M},\overbrace{c_2,\ldots,c_2}^{M},\ldots,\overbrace{c_N,\ldots,c_N}^{M}\right]\right)$$

$$= \frac{1}{MN\sum_{i=1}^{N} Mc_i^p}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M^2(c_j-c_i)^p$$

$$= \frac{1}{N\sum_{i=1}^{N} c_i^p}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(c_j-c_i)^p = S_p(\mathbf{c}).$$

Theorem 8

GDS satisfies **P7: Bill Gates** and **P8: Babies**.

Proof.

The proof is straightforward by combining Theorems 4, 6 and Robin Hood (cf. Appendix) with Theorems 2.1 and 2.2 presented in [Hurley & Rickard, 2009], the former of which states that if Robin Hood and Scaling are satisfied, then Bill Gates is also satisfied and the latter of which states that if Robin Hood, Scaling and Cloning are satisfied, then Babies is also satisfied.

Theorem 9

GDS satisfies **P9: Saturation**.

Proof.

It can be trivially shown that $S_p(\mathbf{0}_N||1) = \frac{N}{N+1}$. Therefore,

$$\frac{S_p(\mathbf{0}_N||1)}{S_p(\mathbf{0}_{N-1}||1)} = \frac{\frac{N}{N+1}}{\frac{N-1}{N}} \xrightarrow{N\to+\infty} 1$$